University of Alberta

Spatiotemporal Modeling of Ambient Sulfur Dioxide Concentrations in Rural Western Canada

By

Shihe Fan ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Science

in

Medical Sciences – Public Health Sciences

Edmonton, Alberta.

Fall 2006

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-22262-1
Our file Notre référence
ISBN: 978-0-494-22262-1

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

To assist the investigation into the health effects of sulfur dioxide (SO_2) pollution, a spatiotemporal model was developed using a discrete process convolution approach for the SO_2 data collected in rural areas of Saskatchewan, Alberta, and British Columbia. The proposed spatiotemporal model was found to be flexible, allowing us to predict the SO_2 exposure at any single time point or over the entire study period for any single locality, or any sub-region, or the entire study region, regardless of whether or not a location within the coverage of the model was actually monitored for SO_2. The potential use of the model in epidemiological studies is demonstrated and future direction of research is discussed in the thesis.

Acknowledgements

With the conclusion of my studies for this Master of Science degree, I would like to take this opportunity to thank my Supervisor, Dr. A. (Sentil) Senthilselvan from the bottom of my heart for his encouragement, guidance, supervision and support. With all of these, he has made the challenges easier for me in conducting my studies and thesis research. I thank Drs. Igor Burstyn and Nicola Cherry for serving on my supervisory committee and providing guidance and supervision. Dr. Burstyn devoted his valuable time and provided computer resources to this research work. I benefited from the working relationship with him. My thanks go to Dr. Stephen Newman for his service on the examination committee.

I thank the Western Interprovincial Scientific Studies Association for providing the SO_2 data. Without its assistance, there would not have been this thesis.

I am heavily indebted with my friend, Dr. Fangliang He, Associate Professor of Biodiversity and Landscape Modeling at the Department of Renewable Resources, University of Alberta. It is his encouragement and support in numerous ways that persuaded me to take on the challenge of studying for this degree. During the years of befriending with and working for him, I have benefited tremendously from him.

Last, but not least, I thank my wife, Jinyu Xiao, and daughters, Xiangning and Catherine, for their unconditional love and support that they have provided to me in my pursuit of studies and scientific research. Without their love and support, there would not have been my scientific career and achievements.

Table of contents

Page

Chapter 1: Introduction 1

Chapter 2. Background and research objectives 3

2.1. Literature review 3

2.1.1. Introduction 3

2.1.2 Origin, fate, and air quality standards of sulfur dioxide pollution 3

2.1.3. Effects of sulfur dioxide pollution on human health 5

2.1.3.1. Clinical evidence of SO_2 absorption 5

2.1.3.2. Mortality 6

2.1.3.3 Morbidity 7

2.1.4. Spatiotemporal modeling of environmental monitoring data 8

2.1.4.1. Multiple random field/time series models 8

2.1.4.2. Geostatistical models 10

2.1.4.3. Kalman filter 11

2.1.4.4. Hierarchical Models 11

2.1.4.5. Kernel mixing 13

2.1.5. Bayesian modeling 14

2.2. Research objectives 16

Chapter 3. Spatiotemporal modeling of ambient sulfur dioxide concentrations in rural Western Canada 18

3.1. Introduction 18

3.2. Methods 20

3.2.1. Data 20

3.2.2. Descriptive statistics and semivariograms 21

3.2.3. Spatiotemporal modeling 24

3.2.4. Smoothed maps of sulfur dioxide 28

3.2.5. Epidemiological application of the model 28

3.2.6. Consolidation of sub-regions through hierarchical cluster analysis 29

3.2.7. Prediction of ambient sulfur dioxide concentrations for spatial points 31

3.3. Results 32
3.3.1. Descriptive statistics and spatial features of the data 32
3.3.2. Semivariograms 32
3.3.2.1. Spatial semivariograms 32
3.3.2.2. Temporal semivariograms 33
3.3.3. Spatiotemporal modeling results 33
3.3.4. Spatiotemporal patterns of ambient sulfur dioxide concentrations in rural Western Canada 35
3.3.5 Epidemiological application of the spatiotemporal model 36
3.3.5.1. Simulated polygons and random points 36
3.3.5.2. Spatial patterns of the predicted average ambient sulfur dioxide in the study region 37
3.3.5.3. Hierarchical classification of the sub-regions according to the predicted ambient SO_2 concentrations 38
3.3.5.4 Use of the spatiotemporal model for point-referenced epidemiological data 40
3.3.5.5. Use of the modeled results for epidemiological study design 40
3.4. Discussions and conclusions 41
Chapter 4. Discussions and future research 63
References 68
Appendix I. Glossary 82
Appendix II. Kalman filter 85
Appendix III. Conditional distributions 87
AIII.1. Preparations 87
AIII.1.1 Basics of Bayesian statistics 87
AIII.1.2 Posterior distribution of normally distributed data with a normal prior 87
AIII.1.3 Gamma, inverse-Gamma, and scaled-inverse-χ^2 distributions 90
AIII.2 Conditional distribution for μ_t, λ_ε and λ_v 92
Appendix IV. Computing and graphing programs 94
AIV.1. WinBUGS codes for the main spatiotemporal model 94

AIV.2. R codes for data preparation and analysis 95
AIV.3. SAS codes 99
Appendix V. Coordinates of the 64 supporting sites used for the spatiotemporal modeling 102
Appendix VI. History series of the spatiotemporal model parameters 103
Appendix VII. SAS modeling results 107

List of tables

Table 2.1. Air quality criteria, objectives, and standards for sulfur dioxide in Canada and other jurisdictions around the world 17

Table 3.1. Summary statistics of the sulfur dioxide data by month from the year of 2001 to the year of 2002.. 44

Table 3.2. Hyperparameters of the priors for the spatiotemporal model.................. 44

Table 3.3. Posterior mean trends (μ_t) of the spatiotemporal model by month. 45

Table 3.4. Precisions of the measurement (λ_ε) and system (λ_υ) equations of the spatiotemporal model.. 45

Table 3.5. Estimates of the latent variables in the spatiotemporal model in each month from June 2001 to May 2002 46

Table 3.6. Means and standard deviations of the predicted ambient SO_2 concentrations for the four classes shown in Figure 3.13 47

Table 3.7. Estimated variances (with 95% confidence interval) in predicted ambient sulfur dioxide concentrations when the original 200 polygons were classified into four classes versus seven classes 48

Table 3.8. Predicted ambient sulfur dioxide concentrations by month from 2001 to 2002 for the two random spatial points shown in Figure 3.14. 48

Table A5.1. Covariance parameter estimates of the SAS model. 109

List of figures

Figure 3.1. Spatial locations of the sulfur dioxide monitoring sites in rural Western Canada 49

Figure 3.2. Sulfur dioxide monitoring sites and spatiotemporal variation in ambient sulfur dioxide concentrations by month..50

Figure 3.3 Spatial semivariograms of ambient sulfur dioxide concentrations by month, where τ^2 is the nugget, σ^2 the partial sill, and ϕ the range in kilometer 51

Figure 3.4. Semivariogram for January 2002, showing the sill, partial sill, nugget, and range 52

Figure 3.5. Temporal semivariograms of ambient sulfur dioxide concentrations for Site 22 (a) and for all 242 sites (b) 53

Figure 3.6. Spatial locations of sulfur dioxide monitoring sites with nominal coordinates (red circles) and supporting sites on an equilateral distance grid for the spatiotemporal modeling of ambient sulfur dioxide concentrations 54

Figure 3.7. Positive linear relationship between the modeled and the observed mean $\ln(SO_2)$ (a) and the residuals of the modeled $\ln(SO_2)$ (b) 55

Figure 3.8. Smoothed mean surface maps of ambient sulfur dioxide concentrations by month from Jane 2001 to May 2002 56

Figure 3.9. Two hundred simulated Thiessen polygons (irregular polygons in black) and 2000 simulated random points (green dots) overlapping with each of the polygons for prediction of sulfur dioxide exposure in each polygon 57

Figure 3.10. Relationship between polygon area and number of points for the 200 simulated polygons.. 58

Figure 3.11. Predicted ambient SO_2 concentrations (in ppb) from the spatiotemporal model averaged over time and space for each of the 200 polygons 59

Figure 3.12. Hierarchical clusters of the 200 polygons, formed according to the sulfur dioxide concentrations predicted from the spatiotemporal model for each of the 12 months. Numbers on the X-axis represent the sequential polygon numbers and the height on the Y-axis represents the value of the criterion for the particular cluster agglomeration 60

Figure 3.13. Clustered classes of the original 200 polygons with different means of predicted ambient sulfur dioxide concentrations........................ 61

Figure 3.14. Two random locations (A and B) whose ambient sulfur dioxide concentrations are to be predicted from the spatiotemporal model 62

Figure A5.1. Positive linear relationship between the SAS modeled and the measured $\ln(SO_2)$ (a) and the residuals of the SAS modeled $\ln(SO_2)$ (b). The red dots in panel (a) represent questionable SO_2 observations in the original dataset 110

Figure A5.2. Linear correlation between the $\ln(SO_2)$ fitted by SAS (SAS fit) and by the spatiotemporal mode (ST fit) 111

List of statistical symbols

μ	Mean
σ	Standard deviation
σ^2	Variance
ε	Error
λ	Precisions in the spatiotemporal model, i.e. the inverse of σ^2
γ	Semivariance (or semivariogram)
τ^2	Nugget, i.e. the intercept in a semivariogram
ϕ	Range, i.e. the distance at which maximum spatial variance occurs in a semivariogram

Chapter 1

Introduction

Sulfur dioxide (SO_2) is a natural and an anthropogenic air pollutant (Lévêque 2003). The average ambient SO_2 concentration in dry, unpolluted air is less than 0.3 $\mu g\ m^{-3}$ (Harrison 1990). In comparison, annual mean SO_2 concentrations in Canadian cities range from approximately 3 $\mu g\ m^{-3}$ in Winnipeg, Regina, and Saskatoon, 8 $\mu g\ m^{-3}$ in Edmonton and Calgary, to 26 $\mu g\ m^{-3}$ in Halifax (WBK & Associates Inc. 2003).

Sulfur dioxide has reportedly been associated with adverse effects on human health (Bernstein et al. 2004, Routledge and Ayres 2005). Its oxidative products, such as aerosols and salts of sulfate (particularly those with diameter $< 1\mu m$) may be even more damaging than the gaseous SO_2 because they can research the lower respiratory system in the lung (Waldbott 1978, Bernstein 2004). These fine particles, like the gaseous SO_2, can be suspended in the air for a prolonged period (1 to 8 days (Katz 1977, Hidy 1994)) and be easily carried far away from their source of production by air currents. One example of such long distance transportation is the acid deposit in Eastern Canada from the United States of America (Environment Canada 1997).

The exact SO_2 concentration in the near ground-level atmosphere is therefore difficult to predict for a given location at a given time. This uncertainty in the spatiotemporal distribution of SO_2 concentrations creates difficulties for epidemiological studies on the health effect of SO_2. As epidemiological studies are mostly observational, they rely on an accurate assessment of the exposure. Errors in exposure measurements can attenuate or severely distort the association between exposure and health outcome, as well as reduce the statistical power of the studies (Armstrong et al. 1993, Armstrong 1998, 2003).

One way to ascertain the human exposure to a continuous variable like SO_2 may be to monitor them continuously for 24 hours a day around the year. However, practicality often leads to the monitoring of the locality of the residence or the working environment of those exposed people as a proxy of the true exposure. In studies covering a large area, a conventional practice is to establish a network of environmental monitoring sites in the study area for data requisition.

This approach, though economical, leaves unmonitored holes between the monitored sites. The air pollution levels in those un-monitored sub-areas need to be interpolated using the data obtained at nearby monitoring sites. One way to conduct this spatial interpolation is to employ spatial statistic models, such as the Kriging method used in *geostatistics* (Cressie 1994), which exploits similarities/dissimilarities in measurement variances and covariances between monitoring sites. When an explicit time factor is incorporated, a spatial model becomes a spatiotemporal model.

The research presented in this thesis describes an application of a spatiotemporal modeling method to analyze ambient SO_2 concentrations that were measured during a twelve month period over a large area in three Western Canadian provinces, i.e., Saskatchewan, Alberta, and British Columbia. The results from the model will provide a basis for assessments of SO_2 exposure in subsequent epidemiological studies.

The thesis is organized in four chapters. Following this introduction chapter, Chapter 2 describes the background and research objectives. Chapter 3 details the data, statistical methods, results, as well as examples of application of the model in epidemiological studies. Chapter 4 covers the discussion and a brief outline of future research.

On the way to developing this thesis, many mathematical formulae and statistical terms are used. To preserve the readability, some mathematic formulae are provided whereas others are omitted. Some spatial statistical terms are italicized following their first use in the text. A more detailed explanation of these terms is then given in Appendix I. This treatment is designed to focus on the delivery of messages, with the assumption that the reader has already possessed some basic knowledge of spatial statistics. For similar considerations, all tables and figures are placed at the end of each respective chapter.

Chapter 2

Background and research objectives

2.1. Literature Review

2.1.1. Introduction

Sulfur dioxide (SO_2) and its derivatives, i.e. sulfur trioxide (SO_3), sulfurous acid (H_2SO_3), and salts of sulfuric acid (H_2SO_4), are air pollutants that have many detrimental effects on the health of human beings, animals, natural ecosystems, and civil engineering structures (National Air Pollution Control Administration 1969). However, this thesis focuses exclusively on spatiotemporal modeling of SO_2 and only a brief literature review of the health effects of SO_2 is included in this section.

Statistical modeling of spatiotemporal phenomena in environmental monitoring data can be achieved through many different approaches. Each of them has its strengths and weakness, being applicable to certain specific situations, but unsuitable to others. The review on spatiotemporal modeling will selectively cover only those methods that are frequently used in modeling air pollution, particularly those that are close in spirit to the methodology used in this thesis research.

2.1.2. Origin, fate, and air quality standards of sulfur dioxide pollution

Sulfur dioxide is a heavy, colorless gas with a pungent odor. Natural sources of SO_2 in the environment include releases from volcanoes, ocean sprays, aquatic and terrestrial microbial activities, and wildfire combustion of biological materials, such as forests, grassland (Komarnisky et al. 2003, Lévêque 2003, Michaud et al. 2005). Anthropogenic sources of SO_2 consist of combustion of solid and liquid fuels that contain sulfur (i.e. wood, coal, and petroleum products), ore smelting and roasting, Kraft and sulfite wood pulping, sour gas processing, and other minor sources (Western Research & Development 1978). The burning of fossil fuels (coal and petroleum products) is the primary source of sulfur pollution in the United States (National Air Pollution Control

Administration 1969). In comparison, upstream oil and gas production is the biggest contributor of SO_2 emissions in Canada, followed by power generation (Environment Canada 2002). Alberta Environment (2005) estimates that half of the total SO_2 emissions of the province come from natural gas processing plants, with oil sands facilities, power plants, gas plant flares, oil refineries, pulp and paper mills and fertilizer plants as other contributors in order of importance.

Once released into the atmosphere, SO_2 can be photochemically oxidized to form H_2SO_3 and H_2SO_4 (Bunce 1994, Clavert and Stockwell 1984):

$$SO_2 + OH^- + (H_2O, O_2, \text{or } N_2) \rightarrow H_2SO_3^- \qquad (2.1)$$

$$HSO_3^- + O_2 \rightarrow SO_3 + H_2O \qquad (2.2)$$

$$SO_3 + H_2O \rightarrow H_2SO_4 \qquad (2.3)$$

where in Equation (2.1), the molecules in the parentheses are carriers that remove excess energy from the reaction. Alternatively, SO_2 can be catalytically oxidized in the presence of H_2O such as in cloud droplets, fog, or on wet surfaces of plants, soil, or water bodies (Friend 1973):

$$SO_3 + \frac{1}{2}O_2 + H_2O \rightarrow 2H^+ + SO_4^{--} \qquad (2.4)$$

Sulfuric acid and other sulfates that formed through the photochemical and catalytic reactions could account for 5% to 20% of the total suspended particulate matters in urban air (National Air Pollution Control Administration 1969).

Sulfur dioxide, along with its oxidized products, can also be removed from the atmosphere through either dry or wet deposition (Lévêque 2003). Dry deposition is the process in which gaseous and particulate species of SO_2 and/or its derivatives are directly collected on land or water surfaces, such as the direct absorption and adsorption of SO_2 by soil or plants. Wet deposition consists of washout and rainout processes (Lévêque

2003, WBK & Associates Inc. 2003). The washout process refers to all removal events that take place within clouds, whereas the rainout process, as so named, means the interception of sulfur-containing particles by falling raindrops and the diffusion-driven uptake of SO_2 by raindrops (Lévêque 2003).

In order to reduce the SO_2 level in the atmosphere, many jurisdictions around the world have established air quality standards (see Table 2.1). The table reveals substantial variation in those standards. In addition to political reasons, the discrepancy in the standards probably stems, to a certain degree, from insufficient scientific understanding of the adverse health effect of SO_2 pollution.

2.1.3. Effects of sulfur dioxide pollution on human health

2.1.3.1. Clinical evidence of SO_2 absorption

Exposure to SO_2 first incites reactions in the upper and lower respiratory tracks. Symptoms include nasal-constriction, broncho-constriction, and stimulated mucus secretion (Koren 1995). These reactions act as a first line of defense to facilitate the absorption of SO_2 in the upper respiratory tract. Experimental studies show that most of the inhaled SO_2 is absorbed between the nose and the pharynx in human subjects (Frank and Speizer 1964, Speizer and Frank 1966, Wolf et al. 1975), but some may reach deeper into the respiratory tract due to conversions to sulfates or other gas-particle chemical reactions. The absorption of SO_2 is much more effective during nasal breathing than oral breathing (95% vs. 70%) (Federal-Provincial Advisory Committee on Air Quality 1987). Nasal removal of SO_2 accounts for 95 to 99% of inhaled SO_2 under resting conditions in both human and animal subjects, but is reduced because the increase in activities and respiratory workload produces a shift from nasal to oronasal breathing (Environmental Criteria and Assessment Office 1982). This points to a potentially greater effect of SO_2 pollution on those who are active outdoors, such as athletes, field workers, and the like, as well as those in the adult population who are hyperresponsive to SO_2 (Nowak et al. 1997), particularly those with asthmatic conditions (Frank 1980).

The 95% – 99% removal efficiency of SO_2 in the upper respiratory tract is obtained at concentrations > 2.62 mg m^{-3} (1 parts per million (ppm)), but not at ambient SO_2 levels (generally less than 0.1 mg m^{-3} (0.038 ppm)) (Environmental Criteria and Assessment Office 1982). For instance, nasal absorption of SO_2 at ≥ 20 ppm is 80 to 90% in animal models, but is substantially reduced at concentrations ≤ 1 ppm, which may allow SO_2 to reach the bronchi (Federal-Provincial Advisory Committee on Air Quality 1987).

Upon inhalation, SO_2 is rapidly absorbed into the secretions lining the respiratory passages; most is transferred into the systemic circulation. However, during inhalation, SO_2 may react with water to form sulfurous acid (H_2SO_3) or be oxidized into trioxide (SO_3) in the respiratory tract. The SO_3 then reacts readily with water to form sulfuric acid (H_2SO_4), which could then form ammonium sulfate ($(NH_4)_2SO_4$) in the presence of ammonia. Sulfurous acid can readily dissociate to form equilibrium of sulfite and bisulfite ions. Bisulfite ions can sulfonate with biological molecules by auto-oxidation or by addition to cytosine. If not utilized by the human body in any metabolic biochemical reaction, the inhaled SO_2 may be detoxified in the liver and other organs through the sulfite-oxidase pathway to form sulfate salts, which are then carried by blood streams to the kidneys, and excreted in urine (Environmental Criteria and Assessment Office 1982).

2.1.3.2. Mortality

Sulfur dioxide pollution is attributed to cause excess deaths by exacerbating conditions in the young and the elder populations as well as patients with predisposed conditions of heart and respiratory system diseases, such as asthma, bronchitis, cystic fibrosis, emphysema, and cardiovascular diseases (Environmental Criteria and Assessment Office 1982). An earlier incident occurred between December 1 and 5, 1930 in the Meuse Valley, Belgium. A blanket of heavy fog, with an estimate of SO_2 and sulfuric acid combined to 25,000 μg m^{-3}, caused several hundreds of illness and 63 deaths (Firket 1936). To stress the severity of the pollution, Firket stated that if the same condition had occurred in London, England, the death toll might have reached 3200. Indeed, an excess mortality of 4,000 cases was observed during the smog with the SO_2 concentration of 1.24 ppm (~ 4500 μg m^{-3}) in December 1952 in the great London area (Wilkins 1954). A marked increase in deaths was noticed in respiratory and cardiovascular deaths, as well as other causes of deaths except deaths to traffic accidents. The National Air Pollution Control Administration (1969), as well as the Environmental Criteria and Assessment Office (1982), cited more incidents of air pollution in several large cities of the United States between the 1950s and 1960s. However, Ferris and colleagues (1980) cautioned that in those earlier episodes of events, smoke in the air, but not SO_2, was probably responsible for the excess mortality and morbidity. They cited Lawther et al. (1970), who reported that after smoke control measures were instilled, episodes of SO_2 pollution above 750 $\mu g\ m^{-3}$ did not produce exacerbations in patients with chronic bronchitis as had been observed before.

Recent studies, however, have consistently reported an increase in mortality due to SO_2 pollution. These include a significant association between SO_2 pollution and natural causes of deaths in nine Italian cities (Biggeri et al. 2005), increased neonatal deaths in Sao Paulo, Brazil (Lin et al.2004), sudden infant death syndrome in 10 Canadian cities (Dales et al. 2004), increased risk of death in diabetes (Kan et al. 2001) and in all causes in Shanghai, China (Kan and Chen 2003). Conversely, a reduction in SO_2 pollution reportedly decreased excess respiratory and cardiovascular deaths (Hedley et al. 2002).

2.1.3.3 Morbidity

The most frequently reported short term effects of SO_2 exposure are increased emergency room visits (Xu et al., 1995, Bernstein 2005, Wilson et al. 2005), worsening of health status in patients with chronic respiratory diseases, such as bronchitis (Lawther et al. 1970, Herbarth et al. 2001), asthma (Chew et al. 1999, Wong et al. 2001, Lin et al. 2003, Barnett et al. 2005), as well as induction of cardiac illness (Martin 1964, Routledge and Ayres 2005). The health effects of SO_2 are particularly acute and severe in young children, such as increased hospital admissions for respiratory diseases, increased history, symptoms, and prevalence of respiratory illness, reduced lung function and airway flows (Timonen KL and Pekkanen 1997, Asgari et al. 1998, Chew et al. 1999, Herbarth et al. 2001, Lin et al. 2003, Barnett et al. 2005). Exposure to SO_2 pollution and other air pollutants reportedly increases adverse pregnancy outcomes (Liu et al. 2003, Sram 2005).

2.1.4. Spatiotemporal modeling of environmental monitoring data

Throughout this thesis, the following notations will be observed: bold large capital letters indicate matrices; bold small capital letters indicate vectors; lower case letters with subscripts indicate individual elements of matrices and vectors; lower case without subscripts indicate scalar; *Italic* letters indicate random variables. The Harrington letters indicate a domain in which a random process is defined.

Consider a random process (y) in space ($\mathcal{S}$) and time ($\mathcal{T}$) as the sum of a space-time (z) and an error (ε) processes:

$$y(s, t) = z(s, t) + \varepsilon(s, t) \quad (\text{for } s = 1, 2, \ldots, N; t=1, 2, \ldots, T) \qquad (2.5)$$

where $\varepsilon(s, t)$ is white noise that is assumed as $\varepsilon(s,t) \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$. The variance of the error term (σ_ε^2) could evolve in time but many treat it as a constant to simplify the estimation.

The space-time process (z) can be further modeled as the sum of a mean trend process (μ) and a zero mean spatiotemporal process (ψ):

$$z(s, t) = \mu(s, t) + \psi(s, t) \qquad (2.6)$$

where $\mu(s, t)$ is explained by covariates that are observed at the same location and time as the observations on the random process (y); $\psi(s, t)$ follows a mean zero process.

Environmental monitoring data with both spatial and temporal trends fit the profile of such a random process. Below is a brief review on some of the spatiotemporal modeling methods that have been used on this type of data.

2.1.4.1. Multiple random field/time series models

Kyriakidis and Journel (1999) characterize geostatistical spatiotemporal modeling into two broad approaches, one being multiple random field/time series models and the other being spatiotemporal random field models. Multiple random field models treat the random process as a set of temporally correlated random fields; time series models treat the random process as a set of spatially correlated time series. An intrinsic difference between the multiple random field models and the time series models is that the random field models use only the past information in modeling and forecasting. On the other hand, the time series models can take into account both the past and the present information after the data have been observed.

Models that employ time series techniques generally apply to situations when the monitoring sites are limited in numbers. Typically, a multivariate time series (called spatial time series) approach is taken in the time series models (Bennett 1979). The advantage of this approach is that many of the techniques developed for time series analysis are readily available, but the drawback is that the results are only applicable to the monitored sites. Further modeling is required for mapping in space. Examples of this approach can be found for modeling the wind speed by Haslett and Raftery (1989) and Roccio (2005), for modeling ozone by Alvo and Dabrowski (2000) and for modeling carbon monoxide by Tonellato (2001).

Models following the multiple random field philosophy are generally built on the Markov random field (MRF) concept (Cressie 1993). In contrast to the time series method, MRF models tackle first the spatial aspect of the random process, with the parameters evolving over time. In MRF models, the monitoring sites in a study region S are specified through a neighborhood system $N = \{N_i, i \in S\}$, where N_i is the set of sites neighboring i, $i \notin N_i$ and $i \in N_j \Leftrightarrow j \in N_i$. A random field X is an MRF on S in relation to a neighborhood system N if and only if

$$P(x) > 0 \qquad \forall x \in X \tag{2.7}$$

$$P(x_i | x_{-\{i\}}) = P\,(x_i | x_{N_i}) \tag{2.8}$$

Simply said, in MRF models, given its neighborhood structure, the observation taken at one site in a sub-area in S is conditionally independent of observations taken at any other sites in other sub-domains in S. The Hammersley-Clifford theorem guarantees a valid joint distribution of the random process at all sites on S through this conditional specification (Besag 1974). Typically, the neighborhood is defined through a first order specification on a grid system, meaning that only the neighbors in the four cardinal directions are considered although a second order specification (diagonal neighbors) could be easily accommodated in a MRF model.

The MRF models are typically used for modeling a random process on lattice (i.e. areal) data, similar to those many applications in disease mapping (Elliot et al. 2000). To apply this method for modeling a continuous spatial process of environmental data, one needs first to divide the study region into finite countable number of sub-regions that contain the monitoring sites. The distribution of the random process that underlies the environmental data in the sub-regions is then specified using conditional distribution. This approach was used by Handcock and Wallis (1994), as well as Lavine and Lozier (1999), to model temperature patterns, by McMillan et al. (2005) to model ozone patterns.

2.1.4.2. Geostatistical models

A geostatistical model is constructed similarly as specified in the beginning of Section 2.1.4. The model is composed of a mean component that models the trend, and a residual component that models the spatiotemporal variation of the random field around that trend. Deterministic or stochastic models can be used to specify the trend models. Deterministic trend models employ periodic functions to account for variation in time and polynomial functions for continuity in space. These functions may be used alone or mixed, depending on what underlies the data. A stochastic model emerges when the coefficients of the models are specified as random variables that follows a certain probability distribution (Kyriakidis and Journel 1999).

Among the geostatistics-based spatiotemporal models, some directly use classic geostatistics such as space-time variograms (Sampson and Guttorp 1992, Guttorp et al. 1994, De Iaco et al. 2002, 2003, Fernéndez-Casal et al. 2003) and covariograms (Bocci and Dabrowski 2002). Others extend the geostatistical techniques, for instance, the 'kriged Kalman filter' and corresponding likelihood-based estimation strategy (Mardia *et al.* 1998) or *Markov Chain Monte Carlo* (MCMC) strategy (Sahu and Mardia 2005), as well as the dimension reduction space-time Kalman filter (Wikle and Cressie 1999).

2.1.4.3. Kalman filter

The Kalman filter (Kalman 1960) is a recursive estimation procedure. Although developed initially for solving engineering problems, it has now a wide range of applications in fields like computer graphics, engineering, aerospace tracking, statistical quality control, short-term forecasting, biological and environmental data modeling. A Kalman filtering procedure initializes itself by using the information obtained at time *t-1* to forecast the state of a random process at time *t* before the state is actually observed. After the state is observed, the information obtained is then used to correct the previous forecast to arrive at a new state, which is then used for the forecast of the state at time *t+1*. This forecast-correction-forecast procedure continues until it reaches the final destination of the state of the random process under study. A detailed description of the linear Kalman filter and the space-time Kalman filter are found in Meinhold and Singpurwalla (1983) and in Cressie and Wikle (2002).

2.1.4.4. Hierarchical Models

In recent years, hierarchical specification is increasingly popular in modeling environmental data (Mateu et al. 2003). Aside from benefiting from increased computer power, this trend is aided by a strong desire to infer on the underlying hidden process from the data, which are often subject to measurement errors. The appeal of the hierarchical model specification lies in its flexibility. It breaks down a large, complex model into much simpler smaller components. Each component can then have its own specifications, for instance, a nonlinear component at a lower hierarchy in an otherwise linear model at a higher hierarchy. A recent book by Banerjee et al. (2004) covers a wide range of issues and techniques in hierarchical modeling.

One version of the hierarchical spatiotemporal model specification is the dynamic linear model (DLM) (Banerjee et al. 2004). This type of models (also called state space models) has their origin in time series analysis and system theory. The term "dynamic" stems from the fact that the time series process change with the passage of time, whereas the "state space" terminology originates from systems theory. One subclass of dynamic models is normal dynamic linear model (NDLM).

A simple case of DLM is the general univariate NDLM, which can be specified as

$$\{\mathbf{F}, \mathbf{G}, \mathrm{V}, \mathbf{W}\}_t = \{\mathbf{F}_t, \mathbf{G}_t, \mathrm{v}_t, \mathbf{W}_t\}$$

for each time t, where $\mathbf{F}_t$ is a known (r x 1) design vector; $\mathbf{G}_t$ a known (r x r) transition matrix; v_t a variance scalar; $\mathbf{W}_t$ a known (r x r) error matrix; r the number of components in the parameter vector $\boldsymbol{\theta}_t$.

The quadruple then leads to the following distributions

$$(\mathbf{Y}_t\,|\boldsymbol{\theta}_t) \sim \mathrm{N}(\mathbf{F}_t'\boldsymbol{\theta}_t, \mathrm{v}_t) \qquad (2.9)$$

and

$$(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) \sim \mathrm{N}(\mathbf{G}_t\boldsymbol{\theta}_{t-1}, \mathbf{W}_t) \qquad (2.10)$$

where $\mathbf{Y}_t$ is a vector of observations.

With the initial conditions at time t = 0, Equations 2.9 and 2.10 specify the full definition of a general univariate NDLM as follows:

Measurement equation: $\mathbf{Y}_t = \mathbf{F}_t'\boldsymbol{\theta}_t + \mathbf{v}_t$ $\quad \mathbf{v}_t \sim \mathrm{N}(0, \mathbf{V}_t)$ (2.11)

System equation: $\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t$ $\quad \boldsymbol{\omega}_t \sim \mathrm{N}(0, \mathbf{W}_t)$ (2.12)

Initial information: $(\boldsymbol{\theta}_0|\mathbf{D}_0) \sim \mathrm{N}(\mathbf{M}_0, \mathbf{C}_0)$

Both $\mathbf{v}_t$ and $\boldsymbol{\omega}_t$ are assumed to be internally and mutually independent. $\mathbf{D}_0$ is the initial information set at time t = 0. Equation 2.11 relates $\mathbf{Y}_t$ to $\boldsymbol{\theta}_t$ via a dynamic linear regression with a normal error structure having a time varying observational error variance v_t (which can be specified as constant). Equation 2.12 (also called state equation) defines the one-step Markov Chain of the state vector because of the conditional independence property, i.e. given $\boldsymbol{\theta}_t$, $\mathbf{Y}_t$ is conditionally independent of the other observations at time t-1 and earlier. Equivalently, this is to say that given the present, the future is independent of the past. For time t, $\mathbf{F}_t$ is the design vector of known values of independent variables; $\boldsymbol{\theta}_t$ the state or system vector; $\boldsymbol{\mu}_t = \mathbf{F}_t'\boldsymbol{\theta}_t$ the mean response; $\mathbf{v}_t$ the observational error. For a more spirited discussion on the DLMs, the reader is referred to West and Harrison (1997), as well as Shumway and Stoffer (2000).

Because of their flexibility and hierarchical specification, DLMs allow for incorporation of multiple sources of stochastic processes. As such, they have been widely used in environmental modeling (Shaddick and Wakefield 2002, Huerta et al. 2004) and population ecology (Meyer and Miller 1999, Calder et al. 2003). Many of these applications involve the use of the Kalman filter or its variants.

2.1.4.5. Kernel mixing

Kernel mixing is another way of spatiotemporal model specification. For a long time, kernel mixing has been used for probability density estimation and regression modeling in statistical literature (Silverman 1986). The attractiveness of this method rests with its convenience in introducing non-stationarity while permitting analytic calculation and clear interpretation, a feature that is fully taken advantage of in this thesis research. A further advantage is the ability to handle very large datasets, which is often a challenge to other methods. There are two schools of kernel mixing in spatiotemporal modeling, one attributable to Fuentes (Fuentes 2002a, b), and the other to Higdon (Higdon 1998, 2002, Higdon et al. 1999). Calder et al. (2002) modified the method of Higdon and came up with a discrete process convolution dynamic linear model.

Let $\mathbf{Y}_t = \{y(s, t)\}$ be the measurement vector at the N monitoring sites (s = 1, …, N) in a study region at time t during the monitoring period. A discrete process convolution dynamic linear model without covariates in a matrix format can be written as:

$$\underset{(N \times 1)}{\mathbf{Y}_t} = \underset{(N \times M)(M \times 1)}{\mathbf{K}\,\mathbf{x}_t} + \underset{(N \times 1)}{\boldsymbol{\mu}\mathbf{1}} + \underset{(N \times 1)}{\boldsymbol{\varepsilon}_t} \qquad (2.13)$$

$$\underset{(M \times 1)}{\mathbf{x}_t} = \underset{(M \times 1)}{\mathbf{x}_{t-1}} + \underset{(M \times 1)}{\mathbf{v}_t} \qquad (2.14)$$

where **K** is a bivariate Gaussian kernel. x is a latent process specified as a zero mean Gaussian random walk process at M supporting sites. M is chosen based on the spatial variability of the random process and the desired model resolution. A large M at shorter inter-site distances models more local variability, whereas a small M at longer inter-site distance models better the global trend of the random process (Higdon 2002). μ is the grand mean with **1** (a Nx1 matrix) signifying its invariance in space due to the absence of covariates in the model. $\boldsymbol{\varepsilon}_t$ and $\mathbf{v}_t$ are vectors representing observational and system error vectors, respectively.

This specification of the model is flexible. Not only can the model incorporate easily covariates for the mean (μ) trend, but also by-pass the need for imputing the missing data. The latter feature is realized by adopting different kernels at different time points.

2.1.5. Bayesian modeling

Bayesian statistical modeling differs from the usual frequentist approach in that 1) it explicitly incorporates subjectivity in the model statement through the use of prior distributions and 2) it treats the model parameters as random variables that follow a certain probability distribution. Therefore, its conclusions about a parameter or the unobserved data are stated in terms of probability, which contrasts to the point estimates in the frequentist statistics.

The core of Bayesian statistics is to find a probability model and to infer on the posterior distribution of the model parameters based on the Bayes' rule. Suppose that there is a joint probability distribution of a model parameter θ and an observed data point y, this joint probability can be written as a product of a prior distribution $p(\theta)$ and a sampling distribution $p(y|\theta)$, i.e.

$$p(\theta, y) = p(\theta)p(y|\theta) \tag{2.15}$$

Conditional on the observed data y, the posterior density of θ based on the Bayes' rule is:

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y \mid \theta)}{p(y)} \propto p(\theta)p(y \mid \theta) \tag{2.16}$$

where

$$p(y) = \sum_{\theta} p(\theta)p(y \mid \theta) \tag{2.17}$$

if *y* is a discrete random variable,
or

$$p(y) = \int p(\theta)p(y \mid \theta)d\theta \tag{2.18}$$

if *y* is a continuous random variable.

The posterior distribution is generally obtained through MCMC methods using various sampling methods. One such method is the Gibbs sampler (Gelman and Gelman 1984, Gelfand and Smith 1990, Chen et al. 2000). This is an alternating sampling scheme, which estimates one model parameter a time by conditioning the parameter on the other parameters in the model and the data. For instance, suppose that we have data *y* and a vector of two parameters $(\theta_1,\theta_2)'$ in the model and want to estimate these two parameters through MCMC, we would start the Gibbs sampling procedure in the following steps:

Step 0. Choose an arbitrary starting point $\boldsymbol{\theta}_0 = (\theta_{1,0}, \theta_{2,0})'$ and set i =0.

Step 1. Generate $\boldsymbol{\theta}_{i+1} = (\theta_{1,\, i+1}, \theta_{2,\, i+1})'$ sequentially

$\theta_{1,\, i+1} \sim p(\theta_1 | \theta_{2,\, i}, y)$

$\theta_{2,\, i+1} \sim p(\theta_2 | \theta_{1,\, i}, y)$

Step 2. set i = i +1 and repeat Step 1.

Iterate this procedure n times for n being sufficiently large. After convergence in probability distribution, the samples of $(\theta_1,\theta_2)'$ are from a true stationary posterior distribution (Gelfand and Smith 1990). The distribution quantities of $(\theta_1,\theta_2)'$ can then be estimated from the sample after discarding the iterations during the burn-in period. Details on MCMC using the Gibbs sampler can be found in Gelman and Gelman (1984), Gelfand and Smith (1990), Gelman et al. (1995) and Gilks et al. (1996).

2.2. Research objectives

This thesis research deals with the environmental monitoring data on ambient SO_2 concentrations that were collected over a large land area in Saskatchewan, Alberta, and British Columbia over a period of 12 months. The data contain both spatial and temporal components. To capture both the spatial and temporal information embedded in the SO_2 data, this thesis research takes a spatiotemporal modeling approach to address the following three questions:

1) What is the spatial pattern of the near ground-level SO_2 concentrations in rural Western Canada?
2) How does this spatial pattern evolve over time during the monitoring period?
3) How can the spatiotemporal modeling results be used in epidemiological studies?

Table 2.1. Air quality criteria, objectives, and standards (ppm, parts per million) for sulfur dioxide in Canada and other jurisdictions around the world.*

Agency		SO2 concentration			Average time
		Threshold	μg/m³	ppm	
Canada[†]	National	Acceptable	900	0.34	1-hour
			300	0.11	24-hour
			60	0.02	1-year
		Desirable	450	0.17	1-hour
			150	0.06	24-hour
			30	0.01	1-year
		Tolerable	800	0.30	24-hour
	Alberta		450	0.17	1-hour
			150	0.06	24-hour
			30	0.01	1-year
	B.C.	Level "A"	450	0.17	1-hour
			160	0.06	24-hour
			25	0.01	1-year
		Level "B"	900	0.34	1-hour
			665	0.10	24-hour
			75	0.03	1-year
	Ontario		690	0.25	1-hour
			275	0.10	24-hour
			55	0.02	1-year
	Quebec		1310	0.50	1-hour
			228	0.09	24-hour
			52	0.02	1-year
USA	National	Secondary	1300	0.50	3-hour
		Primary	365	0.13	24-hour
			80	0.03	1-year
	California		1300	0.50	1-hour
	Florida		260	0.10	1-hour
	Missouri		665	0.25	1-hour
	Montana		665	0.25	1-hour
			60	0.02	1-year
	New York		665	0.25	1-hour
	Vermont		260	0.10	1-hour
			60	0.02	1-year
Argentina			75	0.03	30-day
Belgium			150	0.06	1-year
Columbia			75	0.03	1-year
Denmark			750	0.30	30-min.
Finland			750	0.30	30-min.
			75	0.03	1-year
Italy			750	0.30	30-min.
Japan			260	0.10	1-hour
			100	0.04	24-hour
Switzerland			30	0.01	Annual
Germany			750	0.30	30-min.
W.H.O.		Acceptable[I]	60	0.02	1-year
		Desirable[']	40	0.015	1-year

* Taken from Federal-Provincial Advisory Committee on Air Quality (1987)

[†] All other unlisted provinces follow Canada's air quality objectives.

[I] 98% of 1-hour averages < 200 μg/m³.

['] 98% of 1-hour averages <120 μg/m³.

Chapter 3

Spatiotemporal modeling of ambient sulfur dioxide concentrations in rural Western Canada

3.1 Introduction

Sulfur dioxide pollution has been linked to adverse effects on human health (McManus et al. 1989, Koren 1995, Van Burg 1995, Brunekreef and Holgate 2002, Komarnisky et al. 2003). Recent epidemiological studies have found a significant association between exposure to SO_2 pollution and mortality of natural causes (Kan et al. 2001, Kan and Chen 2003, Biggeri et al. 2005), neonatal deaths (Lin et al. 2004) and sudden infant death syndrome (Dales et al. 2004). Conversely, a reduction in SO_2 pollution allegedly moderates excess respiratory and cardiovascular deaths (Hedley et al. 2002). Exposure to SO_2 pollution also reportedly increases morbidity, such as increased hospital emergency room visits (Xu et al. 1995, Wilson et al. 2005), bronchitis (Lawther et al. 1970, Herbarth et al. 2001), asthma (Chew et al. 1999, Wong et al. 2001, Lin et al. 2003, Barnett et al. 2005), cardiac diseases (Martin 1964, Pope 2000, Brunekreef and Holgate 2002, Lee et al. 2003, Brook et al. 2003, Brook et al. 2004).

Western Canada is a major oil and gas producing region. Environmental data have shown that upstream oil and gas production is the primary source of the near ground-level atmospheric SO_2 in Canada (Environment Canada 2002). In Alberta, natural gas processing plants, oil sands facilities, oil refineries, and other industrial activities, such as power generation, are the major contributors to the near ground-level atmospheric SO_2 (Alberta Environment 2005a). Annual average SO_2 concentrations ranged between 1 to 4 parts per billion (ppb) in the Edmonton and Calgary area, and 0.5 to 3.5 ppb in Northern Alberta during the period from 1990 to 2003 (Alberta Environment 2005b). However, these annual averages do not adequately describe the spatial and temporal patterns of ambient SO_2 concentrations in Western Canada. A spatiotemporal model that portrays those patterns would be helpful to epidemiological studies on the health effect of SO_2 pollution in Western Canada.

Spatiotemporal modeling of environmental monitoring data can be accomplished through a broad range of approaches (Kyriakidis and Journel 1999, Mateu et al. 2003). Dynamic modeling is one of such approaches in that the spatiotemporal model is comprised of a measurement (or observation) equation and a system (or state) equation (Banerjee et al. 2004). The measurement equation specifies what has been measured, whereas the system equation represents the unobservable latent state of the random process that is to be modeled. In the system equation, the process variables are often defined as a function of their prior states to allow the spatiotemporal process to evolve over time. A key advantage of this approach is that multiple sources of variability, whether being linear or nonlinear (Carroll et al. 1997), can be easily incorporated into the model. One example is the framework of the kriged Kalman filter (Mardia et al. 1998) that is used to model and forecast ozone in New York City (Sahu and Mardia 2005). To handle large datasets, Wikle and Cressie (1999) proposed a dimension reduction Kalman filter. Other examples of dynamic spatiotemporal modeling are those of Sanso and Guenni (2000) for modeling non-stationary rainfall data in Venezuela, Shaddick and Wakefield (2002) for modeling daily multivariate pollutant data in London, and Huerta et al (2004) for modeling ozone pollution in Mexico City.

Kernel mixing is another method of spatiotemporal modeling. One example of this method is a dimension-reduction process convolution approach proposed by Higdon (1998, 2002), which is particularly advantageous in handling very large datasets. In modeling the North Atlantic temperature field, he used two kernels, one for the spatial process and the other for the temporal evolution of the spatial process (Higdon 1998). In a modification to Higdon's method, Calder (Calder et al. 2002, Calder 2003, 2005) used only one kernel to mix the spatial process variables, which were, in turn, specified as a function of their prior states for the temporal evolution of the spatiotemporal model. In doing so, the model becomes a dynamic linear model. An added advantage of this modification, among other things, is the flexibility in handling temporally *misaligned* spatial data. This is achieved by using a different kernel for each different time period in the spatiotemporal model, avoiding altogether the usual requirement of data imputation (Little and Rubin 2002).

Our overall goal is to use the available SO_2 monitoring data to develop a model of spatiotemporal patterns of SO_2 in rural Western Canada. Specifically, the main objective is to use the spatiotemporal model 1) to decipher the spatial pattern of the ambient SO_2 concentrations in rural Western Canada, 2) to find the temporal evolution of the ambient SO_2 concentration pattern in the study region, and 3) to predict the SO_2 exposure for localities within the study region where SO_2 was not monitored. Lastly, we will illustrate how the resultant model may be used in environmental epidemiology.

This chapter is organized in a manuscript format with five sections, including data and methods, descriptive statistical and spatiotemporal modeling results, epidemiological applications of results from the spatiotemporal model, and finally, discussion and conclusions.

3.2. Methods

3.2.1. Data

A more detailed description on the data, the sampling strategy and the analysis of the air samples can be found in Davies et al. (2006) and Burstyn et al. (2006). Briefly, the monthly environmental monitoring data on ambient SO_2 concentrations were obtained from the Western Interprovincial Scientific Studies Association. They were collected in the three Western Canadian provinces, Saskatchewan, Alberta, and British Columbia during the period from June 2001 to May 2002. The data were initially collected for studies on cattle health. As such, the spatial monitoring sites were located where the study cattle herds were managed or pastured. On those locations, air monitors (see descriptions below) were set 1.5 to 1.8 m above the ground in flat terrains that were away for at least 10m to 100m from roads, farm houses, farm equipment, transportation corridors, immediate vicinity of local oil and gas facilities, and forest edges. The spatial geographic coordinates of the monitoring sites were then converted to nominal coordinates of S_x and S_y in kilometer before being provided to researchers to protect the confidentiality of the real site locations. The spatial locations of the monitoring sites in Western Canada are shown in Figure 3.1, with the nominal coordinates in kilometers by month in Figure 3.2.

Monthly averaged SO_2 concentrations were measured using PASS SO_2 passive monitors manufactured by Maxxam Analytics Inc. (Mississauga, ON), which also analyzed the air samples. The air samples were taken with a filter impregnated with sodium carbonate/sodium bicarbonate and shipped to the lab in containers sealed with Teflon tape for analysis (Farwell et al 1987). The sulfate ion was extracted from the sampling media with a hydrogen peroxide solution in ultra pure distilled/deionized water. Ion chromatography following US EPA method 300.1 was used to determine the sulfate ion concentrations (Tang et al 1997, Sembulak and Kindzierski 1999).

In a random sample of locations, replicate measurements were collected for a given month. Because the replication was random and highly variable from month to month, those replicates were averaged for each of the locations before being used for the modeling in this thesis. This was to speed up the model fitting process.

Unless stated otherwise, all analyses and modeling were done using natural logarithm-transformed SO_2 concentration data [i.e. $\ln(SO_2)$]. The transformation was based on a preliminary inspection of the histogram of the raw data, which suggested an approximate log-normal distribution.

3.2.2. Descriptive statistics and semivariograms

Descriptive statistics were obtained on raw SO_2 data to examine their characteristics before they were transformed for subsequent analysis and modeling. This was done using the 'summary' function in *R*, a free statistical software (R Development Core Team 2006).

The SO_2 data arose from a random spatial process. They were autocorrelated within a certain distance, i.e. ambient SO_2 concentrations measured at two adjacent spatial points were similar. Beyond a certain distance, the similarity was by chance alone. This spatial feature can be explored by semivariogram, a plot of semivariance versus distance.

To compute the semivariance (γ), denote y_s ($s = 1, \ldots, N_t$ and N_t the number of SO_2 monitoring sites in each month) the $\ln(SO_2)$, the (spatial) variance of y_s is then:

$$\mathrm{Var}(y_{s+h} - y_s) = \mathrm{var}(y_{s+h}) + \mathrm{var}(y_s) - 2\,\mathrm{cov}(y_{s+h}, y_s)$$

$$\Rightarrow\ 2\,\gamma(h) = 2\,C(0) - 2\,C(h) \qquad (3.1)$$

where h is the separation distance between two spatial points, called lag. Divide both sides of Equation 3.1 by 2:

$$\gamma(h) = C(0) - C(h) \qquad (3.2)$$

and the result is semivariance.

As can be seen in Equation 3.2, γ is a function of h. Theoretically, $\gamma(0) = 0$ for $h = 0$ (i.e. no variability for a data point in itself), and $\gamma(h) =$ sill for $h = \phi$. Nevertheless, $\gamma(0) = 0$ for $h = 0$ exists mostly in theory, but seldom in practice. There are unresolved, sub-grid local variability and measurement errors embedded in the SO_2 data. Consequently, $\gamma(0) = \tau^2$ (notation as in Figure 3.3) for $h = 0$, where τ^2 is called nugget, a term that originates from geology and mining. It is the intercept on a semivariogram. The sill is the total (or maximum) variance as ϕ goes to infinity, whereas ϕ is the range of spatial autocorrelation of the $\ln(SO_2)$ data. Beyond ϕ, the $\ln(SO_2)$ (equivalently, the original SO_2) data points are no longer spatially autocorrelated. The difference between the sill and the nugget is called partial sill (σ^2). These concepts are graphically explained in Figure 3.4.

Semivariograms computed from sampling data are called empirical (or experimental or sampling) semivariograms. Because sampling data contain measurement errors, the resultant empirical semivariograms are often jiggled and are therefore rarely used in spatial modeling. Instead, a theoretical semivariogram model (such as a Gaussian, or a spherical, or a Matern semivariogram) is often fitted to the empirical semivariogram to find the range, (partial) sill, and nugget (see Figure 3.4).

In this thesis, omnidirectional semivariance was computed on the residuals after the spatial trend of the ln(SO_2) data was removed with a second order polynomial, using a binning method with 25 km lags in the geoR package (Ribeiro and Diggle 2001) in R. Semivariance is called omnidirectional because it is computed from all spatial point pairs that fall into a spatial distance class without consideration for the directional relationship between the point pairs. (This is in contrast to the directional semivariance, which takes not only distance, but also direction into consideration when binning the spatial pairs into a class). To implement the computation, all SO_2 monitoring sites were first paired and then binned according to their separation (Euclidean) distance into distance classes and the number of such pairs in each bin recorded. The empirical *semivariance* ($\hat{\gamma}_{ij}$) for the (i, j)th bin was then calculated according to Ecker and Gelfand (1999):

$$\hat{\gamma}^{*}_{ij} = \frac{1}{2N_{B_{ij}}} \sum_{\{(k,l):(s_k - s_l)\in B_{ij}\}} [y(s_k) - y(s_l)]^2 \qquad (3.3)$$

where N_{Bij} was the number of sites in bin B_{ij} and $y(s_.)$ was the ln(SO_2).

A Matern correlation function of the form

$$\rho(h) = \frac{1}{2^{(\kappa-1)\Gamma(\kappa)}} (h/\phi)^{\kappa} K_{\kappa}(h/\phi) \qquad (3.4a)$$

was fitted to the empirical semivariograms to determine the range, partial sill and nugget for each month. In Equation 3.4a, h is the distance between pairs of monitoring sites; $\phi > 0$ is a scale parameter; $\kappa > 0$ (varied between 0.4 and 0.45) is a shape parameter that controls the differentiability of the underlying random process; K_κ is a modified Bessel function of the third kind of order κ; $\Gamma(\kappa)$ is the usual Gamma function.

Occasionally, the Matern function did not fit the empirical semivariogram well. Under such circumstances, a spherical correlation function in the following form was used:

$$\rho(h) = \begin{cases} \theta_0 + \theta_1\left(\frac{3}{2}\frac{h}{\phi} - \frac{1}{2}\left(\frac{h}{\phi}\right)^3\right) & 0 \le h \le \phi \\ \theta_0 + \theta_1 & h > \phi \end{cases} \quad (3.4b)$$

where θ_0 (τ^2 in Figure 3.3), θ_1(σ^2 in Figure 3.3.), and ϕ are respectively the nugget, partial sill and range (see Figure 3.4)

Temporal semivariograms of SO_2 measurements were fitted to data obtained at some single sites as well as on a subset of sites that were monitored for at least 11 months (n = 242). This was done using a SAS program shown in Appendix IV.

3.2.3. Spatiotemporal modeling

The spatiotemporal modeling followed broadly a dynamic process convolution model specification (Calder et al 2001, Calder 2003, Calder et al. 2003, Calder 2005), which was a variation of the process convolution approach used by Higdon (1998, 2002) as described in Section 2.4. Briefly, the $\ln(SO_2)$ [i.e. $y(s, t)$; $s = 1, 2, \ldots, N_t$; $t = 1, 2, \ldots, 12$; s is the s^{th} monitoring site. N_t is the number of sites monitored in month t (see Table 3.1)] was modeled by convolving a zero mean Gaussian random process *X* at each supporting site ω_j for $j = 1, \ldots, M$ (= 64) with a symmetric bivariate Gaussian kernel ($K(s, \omega_j)$). The kernel had a zero mean and standard deviation of 200 km. The standard deviation equaled approximately to the mean range of the monthly semivariograms as described above. The supporting sites, $\omega_1, \omega_2, \ldots \omega_{64}$, were systematically laid on an equilateral grid that covered the entire study region as well as a buffering boundary (Kern 2000) that equals to half of the standard deviation of the Gaussian kernel (Figure 3.6). The starting point for these grid points was the most south-western coordinates of the monitoring sites minus half of the standard deviation of the Gaussian kernel. Thereafter, those grid points that were far away from the study region were manually removed for the lack of data support. The coordinates of those remaining 64 supporting sites are shown in Appendix V. Each process variable $x(\omega_j; j = 1, \ldots, 64)$ followed a Gaussian random walk for $t = 1, \ldots 12$.

Under such a setup, the spatiotemporal model is written as:

$$y(s,t) = \mu_t + \sum_{j=1}^{M} k_t(\omega_j - s)x(\omega_j, t) + \varepsilon_{s,t} \qquad (3.5a)$$

$$x(\omega_j, t) = x(\omega_j, t-1) + \nu_{j,t} \qquad \forall i \in M \qquad (3.5b)$$

where both the $\varepsilon_{s,t}$ and the $\nu_{j,t}$ are mutually independent and normally distributed with zero means. μ_t is a time-varying mean shift suggested by data shown in Table 3.1.

The mean zero bivariate Gaussian kernel is centered on each monitoring site:

$$K_t(d) \propto \exp\left(-\frac{\| d \|^2}{2\sigma^2}\right) \qquad (3.6)$$

where σ equals 200 km, $\|d\| = \|\omega_j - s\|$ is the Euclidean distance between the supporting and the monitoring sites. Note that the kernel varied with time, because the number of monitoring sites differed from month to month, which had caused spatial misalignment in the data. The fewest number of monitoring sites was observed in March 2002 (303) and the highest number of sites was observed in August 2001 (928) (Table 3.1). The use of a separate kernel for each month avoided the need for imputation of missing data as usually practiced under such circumstances (Little and Rubin 2002). From the specifications in Equation 3.5, one can easily derive the correlogram, which is simply the kernel convolving with itself (Kern 2000).

For brevity, Equation 3.5 can be rewritten in a matrix form:

$$\mathbf{Y}_t = \mathbf{K}_t\mathbf{x}_t + \mu_t\mathbf{1} + \boldsymbol{\varepsilon}_t \qquad \boldsymbol{\varepsilon}_t \sim N(0, \lambda_\varepsilon I) \qquad (3.7a)$$

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{v}_t \qquad \mathbf{v}_t \sim N(0, \lambda_\nu I) \qquad (3.7b)$$

where at time t, $\mathbf{Y}_t$ is a length N_t-vector containing $\ln(SO_2)$; $\mathbf{K}_t$ an N_t x M kernel matrix; $\mathbf{x}_t$ a length M-vector of latent system variables; **1** is a length N_t-vector of ones; $\boldsymbol{\varepsilon}_t$ a length N_t-vector of measurement errors; $\mathbf{v}_t$ a length M-vector of system errors. Although being time-varying, μ_t is a spatial constant at time t.

The kernel $\mathbf{K}_t$ in Equation 3.7a was rescaled to ensure the validity of the correlogram (Kern 2000):

$$\sum_{j=1}^{M} k_t^{\,2}(\omega_j - s) = 1 \qquad (3.8)$$

According to the specifications in Equation 3.7a, the kernel $\mathbf{K}_t$ mixes the Gaussian processes $x(\omega_j;\ j = 1, \ldots, 64)$ to model the spatial pattern of the SO_2 data at time t. The $\mathbf{x}_t$ as a function of $\mathbf{x}_{t-1}$ in Equation 3.7b fits the temporal evolution of the spatial random process. Together, they complete the spatiotemporal process for the ambient SO_2 concentration data.

Under the Bayesian paradigm of hierarchical modeling, the joint posterior distribution of all parameters in Equation 3.5 or 3.7 is

$$p(x,\mu,\lambda_\varepsilon,\lambda_v \mid Y) \propto \prod_{t=1}^{T} p(Y_t \mid x_t,\mu,\lambda_\varepsilon)\,p(x_t \mid x_{t-1},\lambda_v)\,p(\mu)\,p(\lambda_\varepsilon)\,p(\lambda_v)\,p(x_0) \qquad (3.9)$$

where $\boldsymbol{\mu} = \{\mu_t;\ t=1, \ldots, 12)$. Distributions of posteriors were obtained through MCMC using a Gibbs sampler (German and German 1984, Gelfand and Smith 1990) in *WinBUGS*, a free Bayesian statistical computing software (version 1.4.1., Spiegelhalter et al. 2003). Two chains were run simultaneously. Convergence was considered achieved if the time history of the parameters in the two chains overlapped each other and stayed flat as shown in Appendix VI, although a more formal analysis, such as the 'coda' package in R (Plummer et al. 2006), could be used for this purpose. The parameters were obtained from 12000 iterations, following 8000 burn-in iterations.

The independent and conjugate priors for the parameters in Equation 3.5 or 3.7 were specified as follows:

$\mu_t \sim N(m_{\mu t}, c_{\mu t})$ (3.10)

$X_0 \sim N(m_0 1, c_0 I)$ (3.11)

$\lambda_\varepsilon \sim$ Inverse-Gamma$(\eta_\varepsilon, \delta_\varepsilon)$ (3.12)

$\lambda_v \sim$ Inverse-Gamma(η_v, δ_v) (3.13)

where $m_{\mu t}$, $c_{\mu t}$, m_0, c_0, η_ε, δ_ε, η_v, and δ_v are all hyperparameters of the priors as listed in Table 3.2.

The WinBUGS program has a built-in expert system to determine automatically the full conditionals for the model parameters according to the prior specifications. Because these priors are all proper conjugates, the full conditional distribution for μ_t, λ_ε and λ_v can all be found in closed form. For completeness, the full conditionals for μ_t, λ_ε and λ_v were provided below (Calder 2003); detailed derivation of these conditionals could be found in Appendix III.

To find μ_t, define

$$a = \sum_{i=1}^{N_t}(y_{s,t} - k_t(s,.)x_t) \tag{3.14}$$

where $K_t(s,\cdot)$ is the s^{th} row of the kernel matrix $\mathbf{K}_t$. The full conditional distribution of μ_t is

$$\mu_t \mid - \sim N\left(\frac{\frac{a}{\lambda_\varepsilon}+\frac{m_{\mu t}}{c_{\mu_t}}}{\frac{N_t}{\lambda_\varepsilon}+\frac{1}{c_{\mu t}}}, \frac{N_t}{\lambda_\varepsilon}+\frac{1}{c_{\mu t}}\right) \tag{3.15}$$

To find λ_ε, we first define

$$SS_\varepsilon = \sum_{t=1}^{T}(\mathbf{Y}_t - \mu\mathbf{1} - \mathbf{K}x_t)'(\mathbf{Y}_t - \mu\mathbf{1} - \mathbf{K}x_t) \tag{3.16}$$

The full conditional distribution of λ_ε is

$$\lambda_\varepsilon \mid . \sim Inverse\_Gamma\left(\frac{NT}{2}+\gamma_\varepsilon, \frac{SS_\varepsilon}{2}+\delta_\varepsilon\right) \tag{3.17}$$

Similarly, we define

$$SS_v = \sum_{t=1}^{T} (\mathbf{x}_t - \mathbf{x}_{t-1})'(\mathbf{x}_t - \mathbf{x}_{t-1}) \tag{3.18}$$

The full conditional distribution of λ_v is

$$\lambda_v \,|\, . \sim Inverse\_Gamma\left(\frac{MT}{2} + \gamma_v, \frac{SS_v}{2} + \delta_v\right) \tag{3.19}$$

The full conditional distribution of the $\mathbf{x}_t$'s was found iteratively following the Kalman filter approach (see Appendix II for details) (Kalman 1963, Meinhold and Singpurwalla 1983).

3.2.4. Smoothed maps of sulfur dioxide

To visualize the modeled results, 2000 random points were simulated using a *simple sequential inhibition process* (rSSI) in the 'spatstat' package of R (Baddeley and Turner 2005). The inhibition distance was arbitrarily chosen at 10 km. It was understood that this distance was unnecessarily short because the spatial autocorrelation of the SO_2 data ranged much farther. The choice of this distance was solely for the purpose of a better interpolation of the modeled surface. The $\mathbf{K_t}$ matrix in Equation 3.7 was recalculated accordingly based on the coordinates of these simulated points in relation to the supporting sites ω_j (j = 1, 2,..., 64). The $\ln(SO_2)$ was then predicted for these spatial points over time using the parameters in Equation 3.7 that were estimated by WinBUGS. After converting back to the original linear scale, the predicted SO_2 surface was interpolated in ArcGIS 9.1 (Environmental Systems Research Institute, Inc. (ESRI) 2005) with an Ordinary Kriging method in the spatial analyst extension.

3.2.5. Epidemiological application of the model

To demonstrate the utility of the spatiotemporal model (Equation 3.7) in epidemiological studies, the study region was defined first by dissolving a 100 km circular buffer surrounding each monitoring site in ArcView GIS3.2 (ESRI. 1999). The boundary of the study region was then defined by joining the outmost edges outlining the buffers (Figure 3.9).

Within the boundary of the study region, 200 random points with an inhibition radius of 50 km were simulated using the rSSI function in the 'spatstat' package of R. *Thiessen polygons* were then created to define the influence area surrounding each of these points in ArcView GIS 3.2. This resulted in 200 Thiessen polygons to cover the entire study region (Figure 3.9). Each polygon symbolizes one areal unit in a hypothetical epidemiological study.

The 2000 random points that were simulated for making the smoothed maps were then superimposed on the Thiessen polygon theme layer in ArcView GIS3.2. The points that intersected with each of the 200 Thiessen polygons were assigned to that polygon (Figure 3.9). The predicted values for all the points that fell within each polygon were averaged over space to arrive at a geometric mean (or arithmetic means) of the predicted ambient SO_2 concentrations for that polygon by month. This created a 200 (polygons) x 12 (months) attribute matrix, which was then used in the hierarchical analysis as discussed below.

A grand geometric mean was also calculated for each polygon. This grand geometric mean (μ_g) is defined as:

$$\mu_g = \exp\left(\frac{1}{N_j T}\sum_{t=1}^{T}\sum_{i=1}^{N_j} y_{it}\right) \tag{3.20}$$

where T = 12, the total number of time points in month; N_j is the number of random points in the j^{th} polygon; y_{it} is the i^{th} $\ln(SO_2)$ at time t that is predicted from the spatiotemporal model (Equation 3.7). This result was used to produce Figure 3.11.

3.2.6. Consolidation of sub-regions through hierarchical cluster analysis

The predicted SO_2 exposure obtained in the step above was continuous, which could be difficult to use in epidemiological studies. The subtlety rests with the precision of the prediction, the magnitude of the error, etc (Tielemans et al. 1998, Armstrong 2003, Kim et al. 2006). Variation in exposure estimates could introduce substantial errors into the final results, making them less reliable. In light of this fact, some non-overlapping parsimonious classes of exposure would be much more desirable. This was achieved using a hierarchical cluster analysis (Romesburg 1984, Timm 2002) of the attribute matrix developed in Section 3.2.5.

To begin with, a dissimilarity matrix was constructed from the Euclidean distance (e_{jl}) of the predicted SO_2 concentrations between the polygons. The matrix was then employed to cluster the polygons hierarchically based on Ward's method (Ward 1963). The Euclidean distance is defined as:

$$e_{jl} = [\sum_{t=1}^{12} (y_{tj} - y_{tl})^2]^{1/2} \tag{3.21}$$

where y_{ij} and y_{il} are the predicted SO_2 concentrations for the jth and lth areal units at time t.

The Ward's method originates from the analysis of variance. It aims to minimize the intra-cluster variance while maximizing the inter-cluster variance. Therefore, the clustering process seeks out the least amount of increase in the sum of squared deviations from cluster means. The error sum of squares (ESS) between each pair of polygons is calculated as:

$$ESS = \sum_{t}^{12} (\sum_{i}^{n} y_{ti}^2 - \frac{1}{n} (\sum_{i}^{n} y_{ti})^2) \tag{3.22}$$

where the first summation is with respect to the predicted SO_2 average for each polygon in each of the 12 months arranged in columns, and n the number of polygons within each cluster. The criterion for joining polygons is that it should produce the smallest possible increase in the error sum of squares.

The hierarchical tree clustering process begins with each polygon as a distinct cluster by itself. Then, the algorithm proceeds iteratively, joining the two most similar clusters together until eventually all polygons are joined together to form only one large cluster in the final step. This iterative process produces a dendrogram with branches of various numbers of polygons joining at different heights. When the dendrogram is cut at a chosen height, the polygons that are linked together into clusters become distinctive classes, achieving the objective of classification. The cluster analysis was performed using the 'hclust' function, and the cut of the dendrogram used the 'cutree' function in the base package of R. The output of the cluster analysis was then re-mapped using ArcView GIS3.2, as shown in Figure 3.13.

Where to cut the hierarchical dendrogram requires the consideration of many factors. Statistically, one may look into the variability or homogeneity within and between classes that the cut results, given the option that the dendrogram can be cut at different heights. One example is to use a linear random effect model to compare the effect of cuts at different heights:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{k(ij)} \qquad (3.23)$$

where y_{ijk} is the $\ln(SO_2)$ value at k^{th} time of j^{th} polygon in i^{th} class and

$$\alpha_i \sim N(0, \sigma_c^2),\ \beta_{j(i)} \sim N(0, \sigma_p^2),\ \varepsilon_{k(ij)} \sim N(0, \sigma_e^2)$$

where α_i is random effect of classes with σ_c^2 as between class variance; $\beta_{j(i)}$ is random effect of polygons (within class) with σ_p^2 as common between-polygon variance; $\varepsilon_{k(ij)}$ is random effect of monthly prediction of $\ln(SO_2)$ within polygon within class with σ_e^2 as common between-month variance. The SAS PROC MIXED codes for this analysis are attached in Appendix IV.

3.2.7. Prediction of ambient sulfur dioxide concentrations for spatial points

The use of the spatiotemporal model (Equation 3.7) for predicting SO_2 exposure at point locations requires only a recalculation of the kernel matrix ($\mathbf{K}_t$). To demonstrate this use, two locations were randomly sampled within the study region. (These points were in the areas where no SO_2 monitoring data were available). The SO_2 concentrations for these two locations were predicted from the spatiotemporal model (Equation 3.7).

3.3. Results

3.3.1. Descriptive statistics and spatial features of the data

The monitored ambient SO_2 concentrations varied widely (Table 3.1), with values being as low as undetectable by the monitoring instrument/methodology (< 0.005 ppb) to as high as 8.35 parts per billion (ppb). The winter months (December 2001 to March 2002) had much higher ambient SO_2 concentrations than the months of the other seasons.

The spatial variability of the SO_2 measurements was evident in Figure 3.2. Two important features were readily identifiable. One was the consistency of the locations with high and low ambient SO_2 concentrations over time. The other was the spatial mosaic of the high and low ambient SO_2 concentrations. Sub-regions with high and low ambient SO_2 concentrations were readily recognizable, for instance, the central high concentration sub-region versus the lower left corner low concentration sub-region in Figure 3.2.

3.3.2. Semivariograms

3.3.2.1. Spatial semivariograms

As described in Section 3.2.2, semivariogram is a convenient display of the spatial variation and the range of the autocorrelation of the $\ln(SO_2)$ data. Shown in Figure 3.3 are the semivariograms of the $\ln(SO_2)$ data for each of the 12 months during the study period. The partial sill of the semivariograms varied between 0.06 (March 2002) and 0.29 (May 2002), with a mean at 0.16 ± 0.07. The nugget fluctuated between 0.03 (March 2002) and 0.22 (October 2001), with a mean at 0.15 ± 0.07. Compared to the partial sill, the nugget was relatively large. The range of the autocorrelation of the $\ln(SO_2)$ data span between 153 (October 2001) and 431 km (May 2002). On average, this range was 217 ± 2.5 km. The above numbers seemingly implied large spatial variation in the $\ln(SO_2)$ data from month to month. However, a careful examination of all the values shown in Figure 3.3 actually suggested otherwise. Both the spatial variation and the range were considerably similar from month to month when the extremes were ignored. This suggested some consistence in the spatial pattern of the ambient SO_2 concentrations over the study region.

Figure 3.4 is the semivariogram for the month of January 2002. Shown in this graph are the sill, partial sill, nugget, and range of the semivariogram. The semivariance increased from 0.1 (nugget) at "zero km" (indicating local variability and measurement error of ln(SO2)) to a maximum of 0.29 (sill or total variance of the ln(SO2)) at 201 km (the range). This left a partial sill of 0.19. This description of the semivariogram for January 2002 can be used to interpret the semivariograms for other months in Figure 3.3.

3.3.2.2. Temporal semivariograms

Shown in Figure 3.5 are the empirical temporal semivariograms for one randomly picked individual site (Sites 22) and for all 242 sites that had been monitored for at least 11 months. The empirical temporal semivariance of the SO_2 data showed a gradual increase starting from 1 time distance (one month) and reached a plateau in 5 time distances before declining again. The smallest semivariance at 1 time distance suggested a strong temporal (auto)correlation at 1 time lag. Indeed, the autocorrelation coefficient was 0.35 for both the single site and all 242 sites averaged. The temporal specification of the random system process (i.e. the **x** variables) in the spatiotemporal model was thus justifiable.

3.3.3. Spatiotemporal modeling results

After 8000 iterations through the WinBUGS program, the model parameters converged (see Appendix VI). The posterior statistics of the spatiotemporal model parameters are listed in Tables 3.3, 3.4 and 3.5, respectively. The results in Table 3.3 indicated that the spatial constant, $\hat{\mu}_{(t;t=1,\ldots,12)}$, did not differ much from zero over time except a higher value in March 2002. This suggested a possibility of simplifying the spatiotemporal model in Equation 3.5 (or Equation 3.7) by removing the time index for this term.

The error precisions for both the measurement equation (λ_ε) and the system equation (λ_υ) were not particularly large, with the system equation error precision slightly smaller than the measurement equation error precision (Table 3.4). So, the system equation error variance (0.23) (which is the inverse of λ_υ) was greater than the measurement equation error variance (0.15) (which is the inverse of λ_ε). A larger spatial variance of the system equation in relation to a smaller error variance of the measurement equation was a desirable effect, because this implied a better spatial predicting power of the model in Equation 3.7. The measurement equation error variance was approximately the same magnitude as the average nugget value on the monthly semivariograms in Figure 3.3. This implied that measurement errors were responsible for much of the nugget effect shown in the monthly semivariograms in Figure 3.3 than was micro variability in the SO_2 data (Waller and Gotway 2004). Replicated measurements seemed to agree with this contention (data not shown).

As expected, the latent variables $\hat{x}(\omega_j, j = 1,...,64)$ showed much variation both in space and in time, with their inter-location spatial variability considerably greater than their intra-location temporal variability (Table 3.5). Because the latent variables $\hat{x}(\omega_j, j = 1,...,64)$ measure the underlying spatial random process, their large inter-location variability is therefore self-explanatory. With the availability of these random process variables, the mathematical spatiotemporal model of Equation 3.5 (or 3.7) is transformed into an analytical model, which can be readily used for spatiotemporal prediction of ambient SO_2 concentrations in the near ground-level atmosphere at any location within the study region at any time point during the study period.

The spatiotemporal model fitted the $\ln(SO_2)$ data well (Figure 3.7).The modeled versus the observed $\ln(SO_2)$ measurements were tightly clustered along a straight line (panel (a)) and the residuals were normally distributed (panel (b)) within the range between -1 and 1, with 86.5% of the residuals being within the range between -0.5 and 0.5. The spatial component of the spatiotemporal model used a moving average approach to estimate the ln(SO_2) for each spatial location. If the data had smooth transition from location to location, this method should mimic well the spatial pattern embedded in the data.

For some reasons, there could be abrupt changes in the spatial data. One such abrupt change was illustrated in red in panel (a) of Figure 3.7. Some of those measurements came with negative values in the original data (39 out of 10295 observations, replicates included) and were treated either as at the detection limit of the instrument/analytical method, or were averaged with a valid value of a replicate if there were replicated measurements at the same time on the same location. As shown in Figure 3.7, neither treatment appeared to be perfect. A better treatment might be either to set the data as missing and let the program to estimate them as unknown parameters through MCMC or to impute the data before modeling them (Hornung and Reed 1990, Little and Rubin 2002, Lubin et al. 2004). One could also choose to delete them from the data set (Rappaport and Kupper 2004), but this would result in a loss of information.

3.3.4. Spatiotemporal patterns of ambient sulfur dioxide concentrations in rural Western Canada

In addition to the estimation and prediction of the ambient SO_2 concentrations for specific localities, an important application of the model is to visualize the spatiotemporal patterns of the ambient SO_2 concentrations in the study region, so that they are easily comprehensible to end-users. Figure 3.8 illustrates the interpolated surfaces of the mean ambient SO_2 concentrations in rural Western Canada in each of the 12 months during the monitoring period. The mean surfaces were interpolated from the SO_2 concentrations that were predicted by the model for the 2000 simulated random spatial points, as explained in Section 3.2.4.

These smoothed maps showed clearly the spatial patterns of the high and low ambient SO_2 concentrations within the study region, and the temporal evolution of those spatial patterns during the study period. Spatially, the three hot spots of high ambient SO_2 concentrations (> 2.0 ppb) occurred respectively in the upper left corner, the upper right corner, and the lower right corner, a high SO_2 concentration sub-region in the middle left, and a low SO_2 concentration sub-region in the lower left corner of the study region were consistently identifiable on the maps across all the 12 months of the monitoring period. These high SO_2 concentration sub-regions all expanded their ranges in space during the winter months from December 2001 to March 2002, and then retreated during the spring of Year 2002. The range expansion of the middle left sub-region with high ambient SO_2 concentrations was the most obvious.

Temporally, the winter months from December 2001 to March 2002 saw the ambient SO_2 concentrations rising to a substantially higher level than those in the months of the other seasons for the entire study region. The exception was in some sub-regions at the edges of the study region, where estimates could be imprecise due to the lack of monitoring sites (Figure 3.8).

3.3.5 Epidemiological application of the spatiotemporal model

3.3.5.1. Simulated polygons and random points

Figure 3.9 shows the 200 polygons symbolizing 200 hypothetical administrative districts within the study region. Each polygon covered a land area ranging from 2793 to 13782 hectares. Figure 3.9 also shows the 2000 simulated random points that were used for predicting ambient SO_2 concentrations for each of the polygons in the study region. As expected, the number of points that intersected with each polygon was variable (Figure 3.9), ranging from 3 to 21 per polygon in proportion to the area of the polygons (Figure 3.10). This is desirable because more predicting points for a large areal unit ensure a better prediction of the areal average of ambient SO_2 concentrations than otherwise. This approach differs from some of the usual practices in spatial statistic literature, which often use only the centroid of the sub-area for prediction (Banerjee et al. 2004).

Although different in approaches, the end results differed little in a simulation like this one (data not shown), because the distance between the centroids of the simulated polygons was much smaller than the range of the spatial variance of the SO_2 data. Points that are too close to each other in space are redundant and contribute little to the statistic power in spatial statistics. However, in a real-world application with large administration districts, the value predicted for the centroid may not be a good representation of the average exposure for the district as a whole. The approach taken in this exercise could be superior (Banerjee et al. 2004).

3.3.5.2. Spatial patterns of the predicted average ambient SO_2 in the study region

The grand geometric mean of the predicted ambient SO_2 concentrations for each of the 200 polygons is shown in Figure 3.11. Several interesting features were apparent in the figure. The first was the 260-fold spatial difference in the averaged ambient SO_2 concentrations (predicted) within the study region, from 0.03 to 7.29 ppb.

The second feature was the striking spatial pattern of the predicted ambient SO_2 concentration averages. Three sub-regions with very high predicted SO_2 concentrations occurred, respectively, in the upper-left, the upper-right, and the lower-right corners of the study region. They extended inward from the edge of the study region for less than 150 km, and being parallel to the edge for only approximately 300 km. A second sub-region with higher predicted SO_2 concentrations occurred in the left-center part of the study region. This sub-region extended roughly parallel to the edges for approximately 1000 km and in perpendicular direction for about 300 km. A third high predicted SO_2 concentration sub-region could be identified at 300 km to the right of the left-center sub-region. This sub-region was approximately 200 km (left-right) x 300 km (up-down). The lower-left corner generally had the lowest average SO_2 concentrations (< 0.4 ppb). Sandwiched between the high and the low SO_2 concentration sub-regions were the sub-regions with intermediate average SO_2 concentrations.

A third feature was that sub-regions with both the lowest and the highest predicted SO_2 concentrations situated at the edges of the study region. A possible explanation for the low SO_2 concentrations in the edge sub-regions could be due to a lack of monitoring sites in these sub-regions, because the study region was extended outward by 100 km (half of the spatial variance range) from the outmost monitoring sites. The highest concentrations occurring at the edges could only be explained by the data, because there were several monitoring sites at those sub-regions that consistently reported high SO_2 readings *all* the time (Figure 3.2).

Comparing Figure 3.11 with Figure 3.8, it appeared that the results observable in Figure 3.11 was a composite of the results shown in Figure 3.8. The predicted SO_2 exposure was categorized into 15 classes in Figure 3.11. In epidemiological studies, these predicted SO_2 concentrations, in ordinal scale, could be used as indicator of SO_2 exposure in any of those sub-regions where health outcomes were measured.

3.3.5.3. Hierarchical classification of the sub-regions according to the predicted ambient SO_2 concentrations

In situations where parsimonious classes of exposure are preferred, the classes shown in Figure 3.11 need to be consolidated. Hierarchical cluster analysis is a quantitative method that is well suited to situations in which one does not have a priori knowledge on how the exposure levels should be classified and would like the data to guide the classification. The hierarchical cluster analysis shown in Figure 3.12 was a data-driven classification method because the number of classes had not been defined a priori. In the clustering process, the computer program placed the more homogenous clusters on the left of the dendrogram, with the heterogeneous clusters on the right. Depending on the research objective (taking into consideration of statistical power, measurement error reduction, etc.), financial feasibility, and personnel resources, the dendrogram could be cut at any height to obtain the desired number of classes. The reduction of measurement errors is particularly worthy of extensive examination when categorizing a continuous exposure variable into discrete classes (Gustafson 2004). Nevertheless, a lower cut results in more classes, lower variability of attributes within each class, but poorer efficiency in the deployment of financial and personnel resources. The opposite is true for a higher cut. For instance, when the dendrogram in Figure 3.12 was cut at the height of eight, seven classes resulted. A cut at 20, on the other hand, reduced the classes to only four, an easily manageable option in a real-world epidemiological study. Shown in Figure 3.13 is the classification of the 200 polygons in Figure 3.11 into four classes, with the mean predicted ambient SO_2 concentrations for each of the four classes listed in Table 3.6.

The information in Figure 3.13 is much easier to read than that in Figure 3.11. All four classes were distinctive from each other. From the lowest to the highest, the predicted ambient SO_2 concentrations doubled at each jump from the first to the third class, but nearly tripled from the third to the fourth class (Figure 3.13 and Table 3.6).

As expected, the variability in the predicted mean ambient SO_2 concentrations also increased as fewer classes were produced. This was illustrated in Table 3.7, in which the variance components of the linear random effect model in Equation 3.21 were demonstrated for both a four- and a seven-class classification. The between-class variance, as well as the standard error of the variance, for the four-class classification was noticeably greater than those for the seven-class classification. The greater standard error for the variance of the four-classes made the hypothesis test on $\sigma_c^2 = 0$ non-significant, in comparison to the statistical significance for the test on $\sigma_c^2 = 0$ for the variance of the seven-classes. Similarly, the variance for the between-polygons (within class) for the four-class classification was more than double of that for the seven-class classification. A large between-class variance is generally preferred in statistical designs, but to what extent one can tolerate the increased within-class variability may need to be balanced with other factors, such as financial and personnel resource requirements, measurement error in exposure assessment, and expected effect size in health outcomes. A large effect size can tolerate greater attenuation due to exposure misclassification (Armstrong et al. 1993, Armstrong 1998, 2003). Effect size is defined as the difference of observed outcomes between two contrasting groups or treatments.

Regardless, the result in Figure 3.13 should be easier to use than that in Figure 3.11 in an epidemiological study. For instance, in a usual logistic regression analysis, one may choose the first level as the reference and let the remaining three groups compare with this reference group. If the exposure to SO_2 had indeed contributed to a human health problem and if the modeled exposure was close to the true exposure of the subjects, the odds ratios for the higher exposure groups would be elevated. If the modeled exposure did not represent the true exposure and if the misclassification of the exposure was non-differential, then, measurement errors in exposure assessment would attenuate the results

(Armstrong et al. 1993, Armstrong 1998, 2003, Kim et al. 2006). On the other hand, if the measurement errors were differential, the results could be either over- or under-estimated (Armstrong et al. 1993, Armstrong 1998, 2003). A safeguard to the exposure ascertainment may be to validate the correspondence between the true and the modeled SO_2 exposure on selected locations and then make adjustments accordingly either in the design or the analysis of the study.

3.3.5.4 Use of the spatiotemporal model for point-referenced epidemiological data

If epidemiological data are point-referenced (e.g. the exact geographic location where subjects reside and receive exposure is known such as in a case-control study), the prediction of exposure to SO_2 from the spatiotemporal model (Equation 3.7) is direct. No additional work is required as compared to the above case for the areal unit data. All that is required is to follow these 4 steps: 1) recalculate the $\mathbf{K}_t$ matrix in Equation 3.4 using the coordinates of both the points and the support sites (ω_i; i=1, …, M) in Appendix V ; 2) multiply the $\mathbf{K}_t$ matrix with the vector of $\hat{x}(\omega_i; i = 1,...,M)$ obtained from the WinBUGS in Table 3.5; 3) add the $\hat{\mu}_t$ constant to the product obtained in step two; 4) obtain anti-logarithm of the results to convert them back to the original linear scale. Table 3.8 demonstrates the monthly results for two such random points, A and B, in the study region (Figure 3.14). Should one wish, a grand mean could be obtained by averaging over time for each of these two spatial points.

3.3.5.5. Use of the modeled results for epidemiological study design

In previous sections, we demonstrated how the spatiotemporal model (Equation 3.7) could be used in epidemiological studies under the assumption that health data had been collected and the results from the spatiotemporal model was used to assign exposure to the study subjects. The results from the spatiotemporal model can also be used to design a study examining the association between ambient SO_2 concentrations and health effects.

In designing such a study, the results from the spatiotemporal model provide information about the hot-spot areas of SO_2 exposure. The researcher would design a study in such a way so that the data-collection (recruitment of cases and controls in a case-control design or assignments of exposure groups in a cohort design) will specifically target more of those hot areas than the other areas. In this way, the researcher can more efficiently use financial and personnel resources. More importantly, s/he may very well improve the research results, because the between-group variation is maximized, while the within-group variation reduced. Consequently, the chances of exposure misclassification are minimized, and the statistical power is increased due to reduced measurement errors in the exposure assessment.

3.4. Discussion and Conclusions

In this chapter, a spatiotemporal modeling approach was used to describe the spatiotemporal patterns of ambient SO_2 concentrations in rural Western Canada. The spatiotemporal patterns revealed by the spatiotemporal model were consistent both in space and time.

The established spatiotemporal model has provided a valuable tool to the design and conduct of epidemiological studies. Its main utility rests with its versatility in predicting ambient SO_2 concentrations in both space and time, whether the required prediction is for a single spatial point location or for an areal unit, i.e. an administration district. Many factors, including the emission sources, the meteorological conditions, the topology of the land mass, and the season of the year, can all influence the spatiotemporal distribution of ambient SO_2 concentrations. The exposure of the at-risk subjects to SO_2 is therefore constantly changing both with time and with spatial location. With the model providing spatiotemporal information on the exposure to ambient SO_2 pollution, the epidemiologist may reduce the rate of misclassifying the exposure by invoking group-based exposure assessment (Tielemans et al, 1998, Armstrong 1998, 2003, Kim et al. 2006).

The smoothed SO_2 maps clearly and consistently identify the high and low SO_2 concentration sub-regions in both space and time. These visualized model results convey sharp contrasts in ambient SO_2 concentrations within the study region at a glance. The user can quickly grasp this spatiotemporal information, as the patterns are almost completely self-explanatory. This information should be helpful to the design of epidemiological studies. For instance, one could strategically sample the sub-regions of interest with the guidance of this information. In such a way, both the financial and the human resources could be more efficiently deployed.

The seasonal variation in ambient SO_2 concentrations shown in the smoothed maps is a well documented phenomenon (Brook et al. 2001), which was also observed in this monitoring program (Burstyn et al. 2006). Monitoring data show that the ambient SO_2 and the fine sulfate particles track each other seasonally in opposite directions. Ambient SO_2 concentrations are high in spring/winter/fall months and low in summer months; sulfate particle concentrations low in spring/winter/fall months and high in summer months. Based on these observations and the physical, chemical principles of SO_2 reactions in the atmosphere, Brook and colleagues (2001) propose that low conversion of SO_2 to particulate sulfate, less in-cloud aqueous-phase oxidation of SO_2 to sulfate, less removal of SO_2 by precipitation, and low air mixing height and strength are the reasons for the higher ambient SO_2 concentrations in winter than in summer months.

A great advantage of this spatiotemporal model is that not only does it provide ambient SO_2 concentration estimation for monitored sites, but also predictions for sites at which there are no monitored data available. The latter case is often more helpful than the former case in an epidemiological study. In this sense, the model extends one's capability of predicting (or assessing) the exposure factor for an epidemiological study. The only pre-caution is that one must evaluate the error of this prediction against the true exposure, perhaps following the methodology advocated by Armstrong (2003). In so doing, one may set up some validation monitoring sites in the study region to take some ambient SO_2 concentration measurements, which are then compared with the model predicted results. If necessary, adjustments are made to reconcile the predicted ambient SO_2 concentrations with the true exposure.

Three scenarios of model uses are demonstrated. One is the use of the model for prediction of SO_2 exposure in areal units. Another is for prediction of ambient SO_2 concentrations at spatial points of interest. The third is for guiding the design of epidemiological studies using the exposure information. These examples reveal the essence and flexibility of the modeling approach. On the way to demonstrating the use of the spatiotemporal model for the areal unit data, a hierarchical cluster analysis method and a variance comparison method were introduced. By employing these methodologies, numerous quantitative classes were consolidated into fewer parsimonious groups. One could then more effectively and efficiently use financial and personnel resources to focus only on those interesting areas where one would expect greater contrast in exposure under a given hypothesis. In so doing, s/he may also improve the validity of the research results due to reduced measurement errors in exposure assessment. These methodologies therefore constitute a valuable part of the spatiotemporal modeling process in epidemiological studies.

Table 3.1. Summary statistics of the SO_2 data by month from the year of 2001 to the year of 2002*.

Month	N_{mea}	N_{site}	Min.	First quantile	Median	Mean	Third quantile	Max.
June 01	1661	890	<det.	0.31	0.45	0.54	0.68	6.40
July 01	1735	927	<det.	0.38	0.53	0.64	0.78	4.60
August 01	1724	928	<det.	0.38	0.54	0.61	0.72	3.43
September 01	1020	910	<det.	0.32	0.47	0.55	0.69	6.58
October 01	895	785	<det.	0.25	0.40	0.49	0.64	8.35
November 01	600	499	<det.	0.39	0.61	0.71	0.89	4.65
December 01	454	361	<det.	0.64	0.94	1.12	1.39	6.00
January 02	430	337	<det.	0.60	1.00	1.13	1.37	7.58
February 02	403	314	<det.	0.53	0.78	0.93	1.11	5.60
March 02	380	303	<det.	0.99	1.40	1.47	1.83	4.30
April 02	389	315	<det.	0.28	0.50	0.61	0.84	3.09
May 02	604	522	<det.	0.23	0.34	0.42	0.48	4.46

*** N_{mea} is the number of individual measurements taken. N_{site} is the number of unique sites monitored for SO_2 in each month. Min. and Max. are, respectively, the minimum and maximum SO_2 concentrations in parts per billion (ppb). <det. stands for below the detection limit (0.005 ppb) of the monitoring instrument/methodology.**

Table 3.2. Hyperparameters of the priors for the spatiotemporal model (Letters in the parentheses correspond to the parameters in Equations 3.8 – 3.11).

	Parameters	
X_0	0 (m_0)	20 (C_0)
μ_t	0 ($m_{\mu t}$)	100 ($c_{\mu t}$)
λ_ε	0.1 (η_ε)	0.1 (δ_ε)
λ_υ	0.1 (η_v)	0.1 (δ_v)

Table 3.3. Posterior mean trends (μ_t) of the spatiotemporal model (in ln(SO_2)) by month.

Time	Mean	SD	Median	Credible interval	
				2.5%	97.5%
June 01	-0.07	0.09	-0.07	-0.25	0.10
July 01	0.07	0.08	0.08	-0.10	0.24
August 01	0.03	0.08	0.03	-0.14	0.20
September 01	0.01	0.09	0.01	-0.16	0.17
October 01	-0.13	0.09	-0.13	-0.30	0.04
November 01	0.05	0.08	0.05	-0.12	0.21
December 01	0.05	0.09	0.05	-0.11	0.22
January 02	0.05	0.08	0.05	-0.11	0.21
February 02	-0.05	0.09	-0.05	-0.22	0.12
March 02	0.22	0.09	0.22	0.05	0.39
April 02	-0.06	0.09	-0.07	-0.23	0.11
May 02	-0.11	0.09	-0.10	-0.29	0.07

Table 3.4. Precisions of the measurement ($\hat{\lambda}_\varepsilon$) and the system ($\hat{\lambda}_\nu$) equations of the spatiotemporal model.

Parameter	Mean	SD	Median	Credible interval	
				2.5%	97.5%
λ_ε	6.58	0.11	6.58	6.35	6.82
λ_υ	4.26	0.84	4.33	3.21	5.74

Table 3.5. Estimates of the latent variables ($\hat{x}(\omega_j, j = 1,\ldots,64)$) in the spatiotemporal model in each month from June 2001 to May 2002.

X	Time											
	2001							2002				
	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
1	-0.72	-0.82	-0.90	-0.88	-1.08	-1.02	-0.69	-0.52	-0.60	-0.70	-0.93	-1.13
2	-0.69	-0.45	-0.44	-0.25	-0.48	-0.38	-0.23	-0.21	-0.41	-0.58	-0.77	-0.85
3	8.76	9.05	9.14	9.43	10.05	10.24	9.88	9.77	9.57	9.48	9.52	9.50
4	-1.95	-1.99	-1.94	-1.63	-1.42	-1.30	-1.41	-1.57	-1.61	-1.74	-1.90	-2.08
5	0.52	0.27	0.29	0.31	0.45	0.47	0.72	0.85	0.82	0.79	0.54	0.38
6	-7.09	-7.18	-7.27	-7.49	-8.26	-8.15	-7.61	-7.42	-7.53	-7.58	-8.00	-7.90
7	1.68	2.01	2.20	2.29	1.60	2.40	2.64	2.67	2.56	2.57	2.36	2.37
8	-7.51	-7.68	-7.80	-8.27	-8.03	-7.38	-7.82	-7.89	-7.98	-8.00	-8.00	-8.16
9	-1.11	-1.17	-1.04	-0.97	-0.64	-0.69	-0.79	-0.84	-0.77	-0.72	-0.82	-0.83
10	10.29	10.32	10.39	10.20	10.55	10.29	10.48	10.55	10.60	10.73	10.43	10.88
11	-1.49	-1.20	-1.10	-1.10	-1.94	-1.57	-1.09	-1.13	-1.15	-0.99	-1.46	-1.01
12	2.03	2.11	2.24	1.90	0.58	1.31	1.50	1.44	1.47	1.67	1.36	1.22
13	0.89	0.55	0.43	-0.12	-0.25	0.14	0.06	0.07	0.15	0.23	0.11	-0.08
14	-8.69	-8.74	-8.89	-9.24	-8.86	-9.18	-9.37	-9.47	-9.46	-9.44	-9.59	-9.43
15	-2.52	-2.28	-2.45	-2.54	-1.96	-2.20	-2.22	-2.43	-2.56	-2.59	-3.03	-2.74
16	-4.36	-4.10	-4.03	-3.75	-4.01	-4.04	-3.77	-3.87	-3.91	-3.82	-4.30	-4.31
17	3.74	3.59	3.65	3.80	3.75	3.69	4.12	4.35	4.63	4.92	4.70	4.62
18	0.79	0.61	0.60	0.59	0.97	1.05	1.30	1.44	1.59	1.68	1.45	1.43
19	1.52	1.55	1.49	1.22	1.29	1.54	1.62	1.63	1.70	1.74	1.65	1.52
20	6.68	6.87	6.76	6.61	6.65	6.90	6.92	6.81	6.67	6.57	6.37	6.22
21	-3.83	-3.86	-3.88	-3.64	-3.33	-3.65	-3.55	-3.52	-3.61	-3.64	-3.84	-3.89
22	-3.15	-3.28	-3.46	-3.50	-3.11	-3.53	-3.40	-3.39	-3.46	-3.47	-3.97	-4.00
23	0.58	0.87	0.91	0.65	0.55	0.44	0.44	0.33	0.19	0.12	-0.36	-0.46
24	-7.97	-7.96	-7.94	-8.09	-8.56	-7.80	-7.57	-7.52	-7.42	-7.53	-7.67	-7.93
25	6.72	6.99	6.99	6.83	6.54	7.07	7.19	7.19	7.18	6.97	7.04	7.20
26	-3.86	-3.86	-3.97	-4.01	-3.92	-4.04	-3.93	-4.15	-4.26	-4.44	-4.69	-4.65
27	3.42	3.77	3.86	3.62	3.52	3.42	3.68	3.64	3.61	3.69	3.17	3.03
28	-4.19	-3.69	-3.33	-3.41	-3.49	-3.36	-3.02	-2.87	-2.75	-2.57	-2.90	-3.00
29	6.28	6.55	6.85	7.32	7.31	7.41	7.54	7.58	7.65	7.48	6.95	6.49
30	-11.36	-11.10	-11.14	-11.05	-11.23	-11.07	-10.82	-10.81	-10.70	-10.96	-11.51	-11.75
31	9.56	9.55	9.44	9.36	9.20	9.39	9.86	9.71	10.02	9.95	9.94	10.07
32	-6.43	-6.64	-6.55	-6.36	-6.15	-6.03	-5.48	-5.67	-5.51	-5.33	-5.41	-5.58
33	4.25	4.39	4.51	4.56	4.70	4.85	5.36	5.54	5.60	5.98	5.64	5.58
34	-3.66	-3.95	-4.08	-4.17	-4.19	-4.16	-3.95	-3.82	-3.63	-3.36	-3.54	-3.59
35	4.77	5.12	5.38	5.72	6.01	5.58	5.57	5.59	5.59	5.58	5.03	4.77
36	-0.63	-0.63	-0.74	-0.94	-0.88	-1.30	-1.17	-1.23	-1.12	-1.00	-1.22	-1.26
37	-2.70	-3.32	-3.24	-3.53	-3.61	-3.78	-3.63	-3.90	-3.94	-3.90	-3.81	-4.18
38	1.66	1.77	1.94	2.19	2.36	2.14	2.15	1.92	1.53	1.63	1.39	1.10
39	-3.21	-3.41	-3.99	-4.28	-4.71	-4.79	-5.02	-5.21	-5.36	-5.13	-5.58	-5.43
40	6.75	6.64	6.69	6.86	6.95	6.82	6.73	6.59	6.50	6.36	6.36	6.07
41	-12.16	-11.92	-11.76	-11.68	-11.41	-11.35	-11.17	-11.20	-11.37	-11.31	-11.43	-11.29

Table 3.5 continued.

42	10.10	10.06	10.14	9.97	10.17	10.33	10.48	10.76	10.77	10.85	11.00	10.98
43	-6.53	-6.60	-6.21	-6.33	-6.12	-6.08	-6.19	-5.94	-6.12	-6.08	-6.09	-6.57
44	9.73	10.09	10.10	10.32	10.25	10.33	10.20	10.23	10.12	10.35	10.10	10.10
45	-11.51	-11.66	-11.81	-11.90	-12.23	-11.79	-11.71	-11.64	-11.48	-11.28	-11.38	-11.34
46	8.10	8.41	8.70	8.93	8.93	9.10	9.48	9.59	9.41	9.12	9.16	8.56
47	3.95	4.51	4.92	4.84	4.50	4.93	5.04	5.11	5.05	4.96	4.78	4.75
48	6.47	6.50	6.52	6.56	6.77	7.08	7.45	7.24	7.14	6.93	6.87	6.82
49	-0.53	-0.77	-1.00	-1.15	-0.85	-0.86	-0.82	-0.58	-0.49	-0.47	-0.43	-0.62
50	-10.17	-10.21	-10.19	-10.17	-10.00	-9.91	-9.88	-9.55	-9.42	-9.23	-9.34	-9.55
51	15.64	15.78	15.87	16.11	16.13	16.62	16.86	17.13	17.43	17.72	17.63	17.65
52	-21.35	-21.50	-21.74	-22.01	-22.34	-21.97	-21.62	-21.40	-21.33	-21.31	-21.34	-21.55
53	4.98	5.20	5.12	5.01	4.83	4.95	5.23	5.33	5.22	5.08	5.04	4.72
54	-2.46	-1.93	-1.77	-2.28	-3.03	-2.86	-3.09	-2.66	-2.48	-2.40	-2.49	-2.50
55	-11.77	-12.08	-11.96	-12.41	-12.41	-12.30	-12.16	-12.29	-12.29	-12.38	-12.48	-12.70
56	1.29	0.80	0.53	0.39	0.89	0.63	0.88	0.40	0.31	0.43	0.30	0.00
57	-2.77	-2.66	-2.94	-2.85	-2.83	-3.13	-3.19	-3.45	-3.64	-3.43	-3.74	-4.09
58	5.42	5.52	5.60	5.45	5.62	5.72	5.58	6.00	6.09	6.06	5.98	5.89
59	4.15	3.78	3.74	3.83	4.32	4.57	4.72	4.83	4.84	5.07	4.85	4.64
60	1.48	1.83	1.69	2.01	1.64	2.00	2.21	2.02	1.83	2.49	1.94	1.58
61	3.58	4.22	4.16	4.46	3.97	4.14	4.26	4.15	3.96	4.34	3.91	3.60
62	1.23	1.20	0.99	0.93	1.47	1.78	1.50	1.80	1.57	1.25	0.96	0.77
63	-4.36	-4.17	-4.36	-4.40	-4.82	-3.98	-4.02	-3.70	-3.79	-3.50	-4.01	-4.16
64	-6.89	-6.38	-6.50	-6.45	-7.41	-6.71	-6.52	-6.37	-6.38	-5.78	-6.21	-6.40

Table 3.6. Mean and standard deviation of the predicted ambient SO_2 concentrations for each of the four classes shown in Figure 3.13.

Group	N	Mean	SD
1	69	0.34	0.20
2	84	0.66	0.38
3	41	1.36	0.93
4	6	3.50	2.77

Table 3.7. Estimated variances in predicted ambient sulfur dioxide concentrations when the original 200 polygons were classified into four classes versus seven classes.

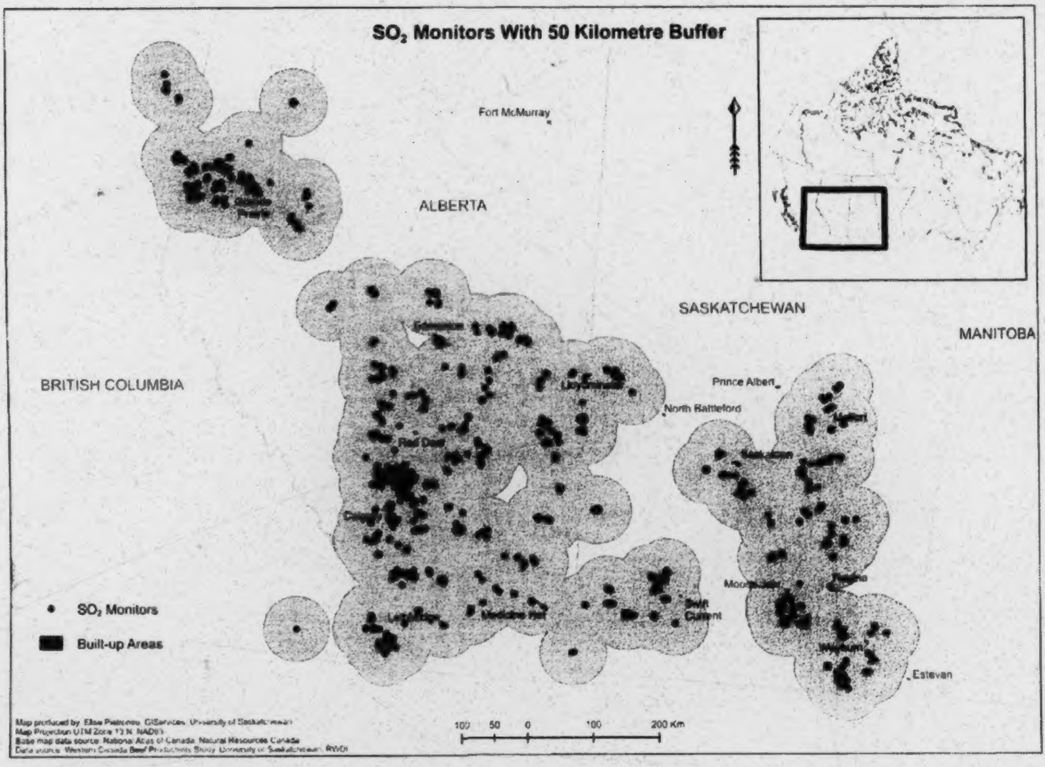
Source of variance	Variance					
	Four classes			Seven classes		
	Estimate	Standard Error	Pr($\hat{Z}$>Z)*	Estimate	Standard Error	Pr($\hat{Z}$>Z)
Between-class	0.9567	0.6844	0.0811	0.9103	0.4898	0.0315
Between-polygon (within-class)	0.0942	0.0114	<0.0001	0.0403	0.0060	<0.0001
Month-to-month (within polygon)	0.2255	0.0068	<0.0001	0.2255	0.0068	<0.0001

* $\hat{Z}$ is a standardized normal statistic.

Table 3.8. Predicted ambient sulfur dioxide concentrations by month from 2001 to 2002 for the two random spatial points shown in Figure 3.14.

Point	Time											
	Jun 01	Jul 01	Aug 01	Sep 01	Oct 01	Nov 01	Dec 01	Jan 02	Feb 02	Mar 02	Apr 02	May 02
A	0.21	0.25	0.28	0.17	0.18	0.35	0.49	0.52	0.43	0.60	0.29	0.18
B	0.45	0.67	0.61	0.56	0.39	0.60	1.16	1.05	0.94	1.29	0.36	0.33

Figure 3.1. Spatial locations of the sulfur dioxide monitoring sites in rural Western Canada.*


SO_2 Monitors With 50 Kilometre Buffer
Fort McMurray
ALBERTA
SASKATCHEWAN
MANITOBA
BRITISH COLUMBIA
Prince Albert
North Battleford
Medicine Hat
Swift Current
Estevan
SO_2 Monitors
Built-up Areas
Map produced by Elise Pietroniro, GIServices, University of Saskatchewan
Map Projection UTM Zone 13 N, NAD83
Base map data source: National Atlas of Canada, Natural Resources Canada
Data source: Western Canada Beef Productivity Study, University of Saskatchewan, RWDI
100
50
0
100
200 Km

* Adopted from Burstyn et al. (2006).

Figure 3.2. Sulfur dioxide monitoring sites and spatiotemporal variation in ambient sulfur dioxide concentrations by month.

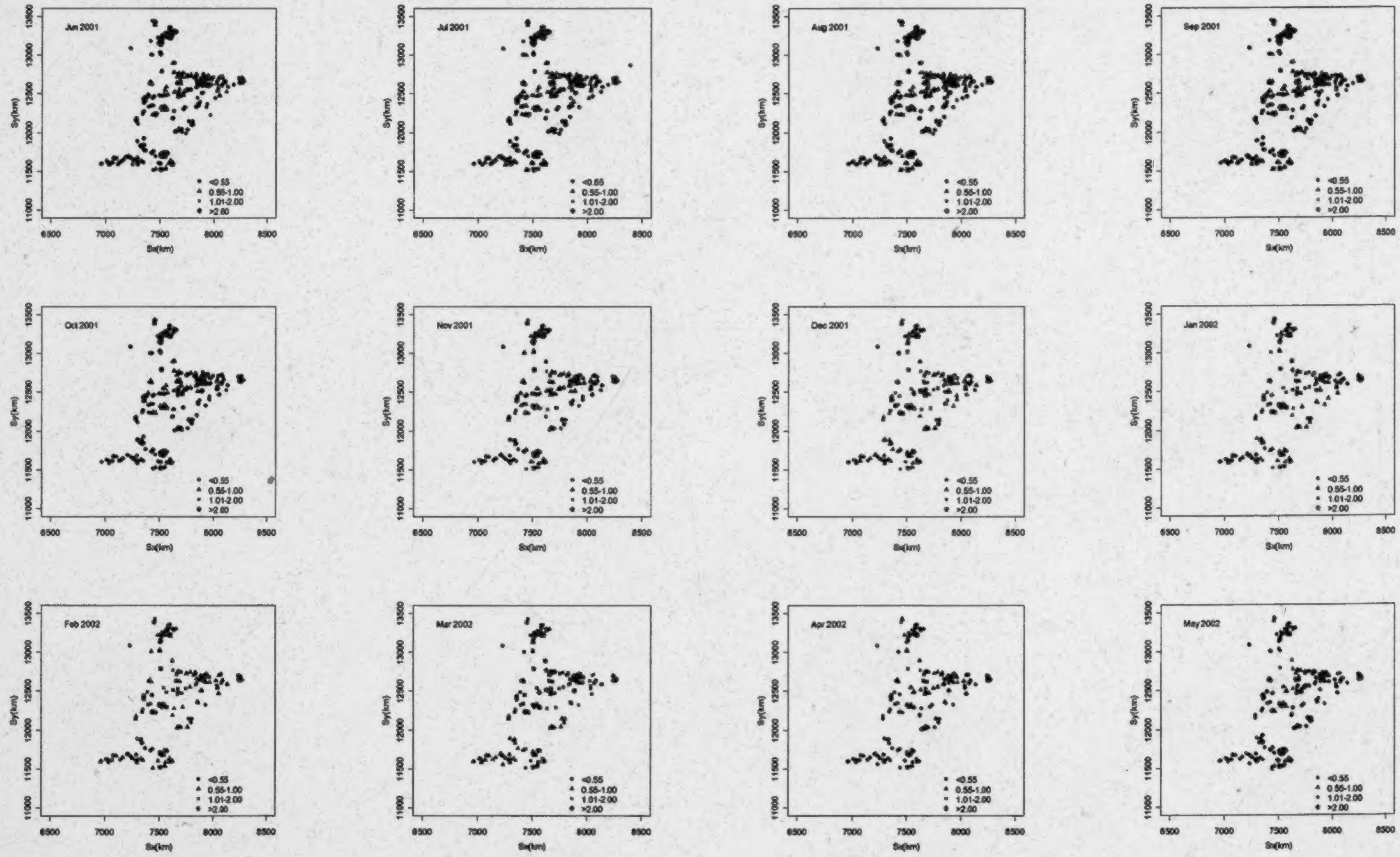


Figure 3.3. Spatial semivariograms of ambient sulfur dioxide concentrations by month, where τ^2 is the nugget, σ^2 the partial sill, and ϕ the range in kilometer.

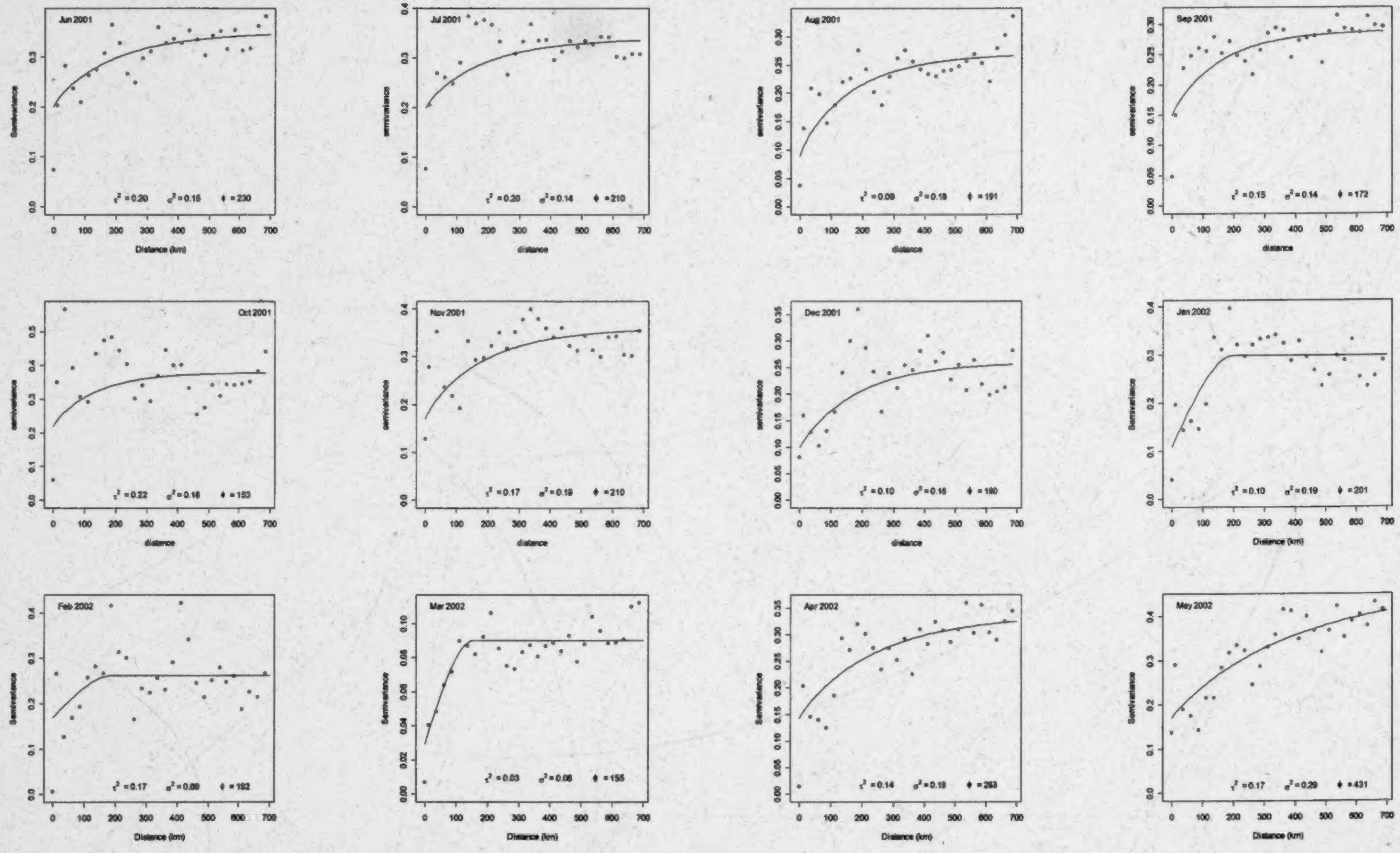


Figure 3.4. Semivariogram for January 2002, showing the sill, partial sill, nugget, and range.

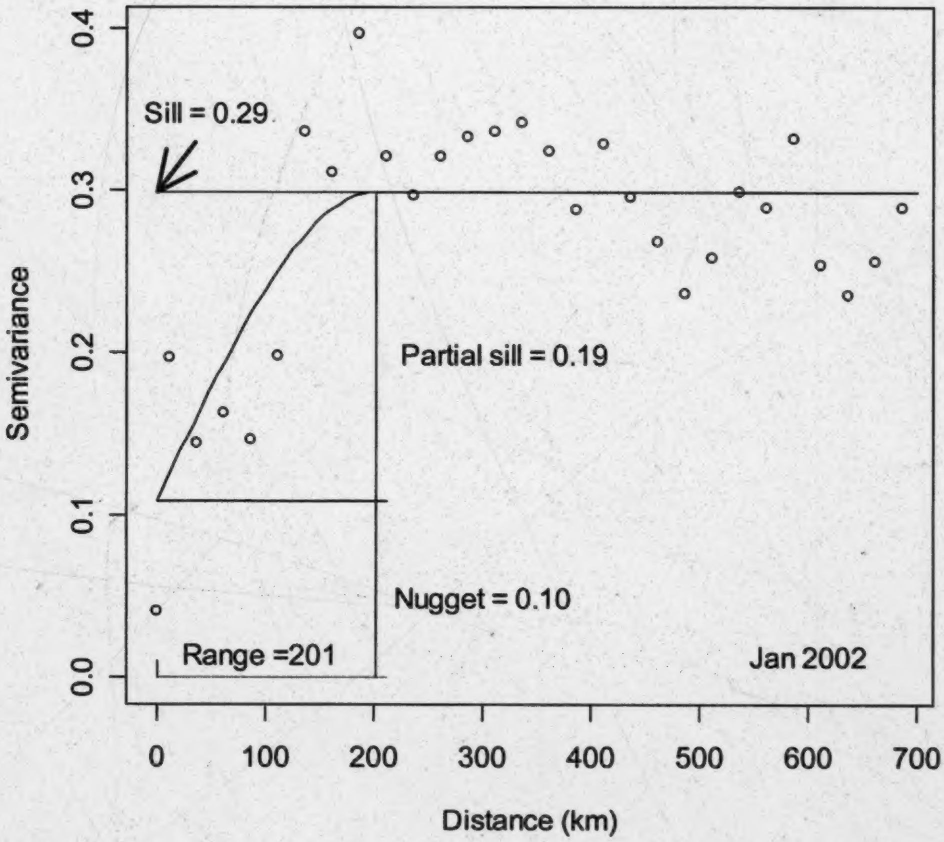


Figure 3.5. Temporal semivariograms of ambient sulfur dioxide concentrations for Site 22 (a) and for all 242 sites (b).

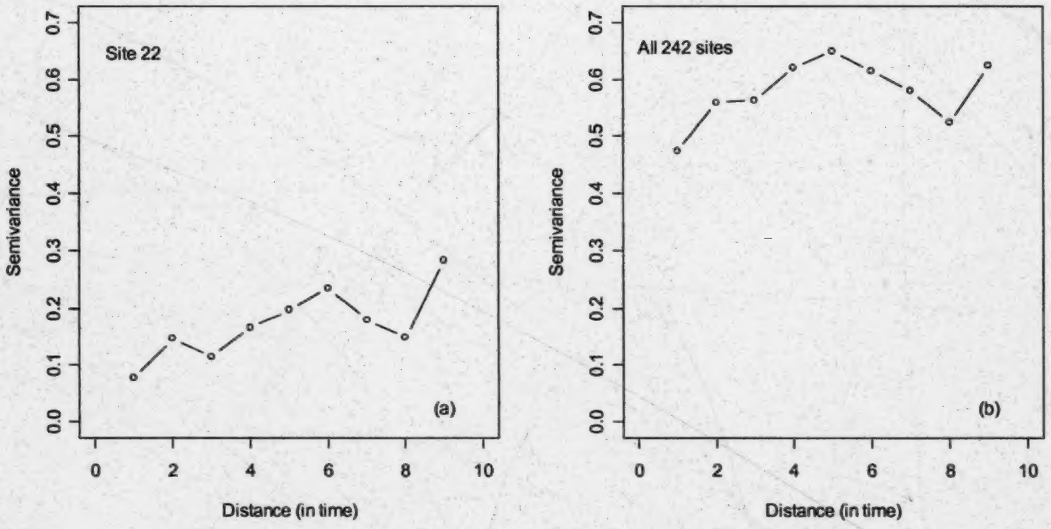


Figure 3.6. Spatial locations of sulfur dioxide monitoring sites with nominal coordinates (red circles) and supporting sites on an equilateral distance grid for the spatiotemporal modeling of ambient sulfur dioxide concentrations.

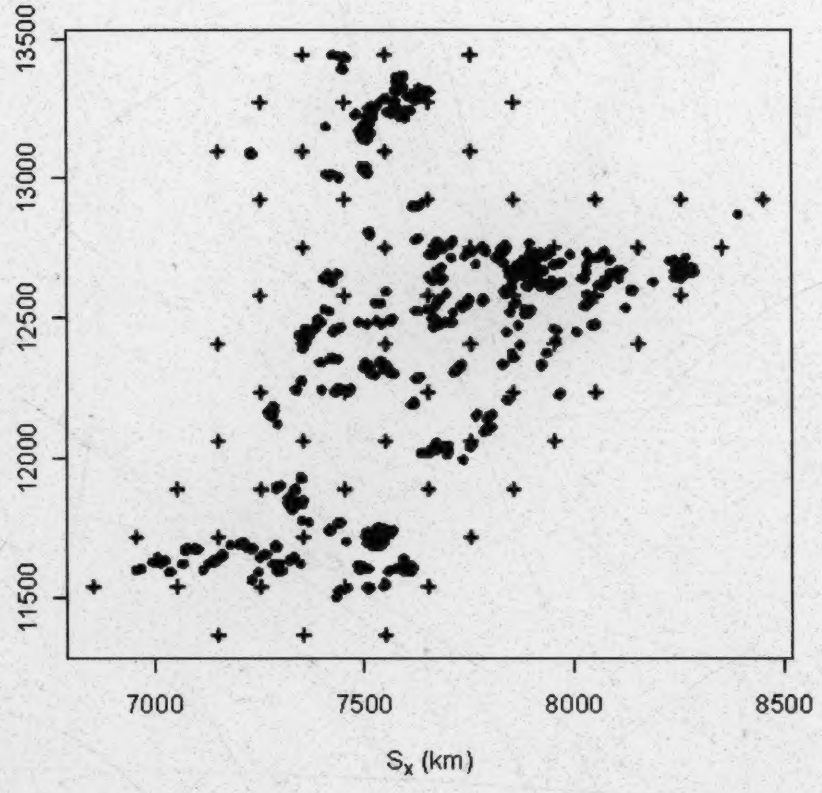


Figure 3.7. Positive linear relationship between the modeled and the observed mean $\ln(SO_2)$ (a) and the residuals of the modeled $\ln(SO_2)$ (b).

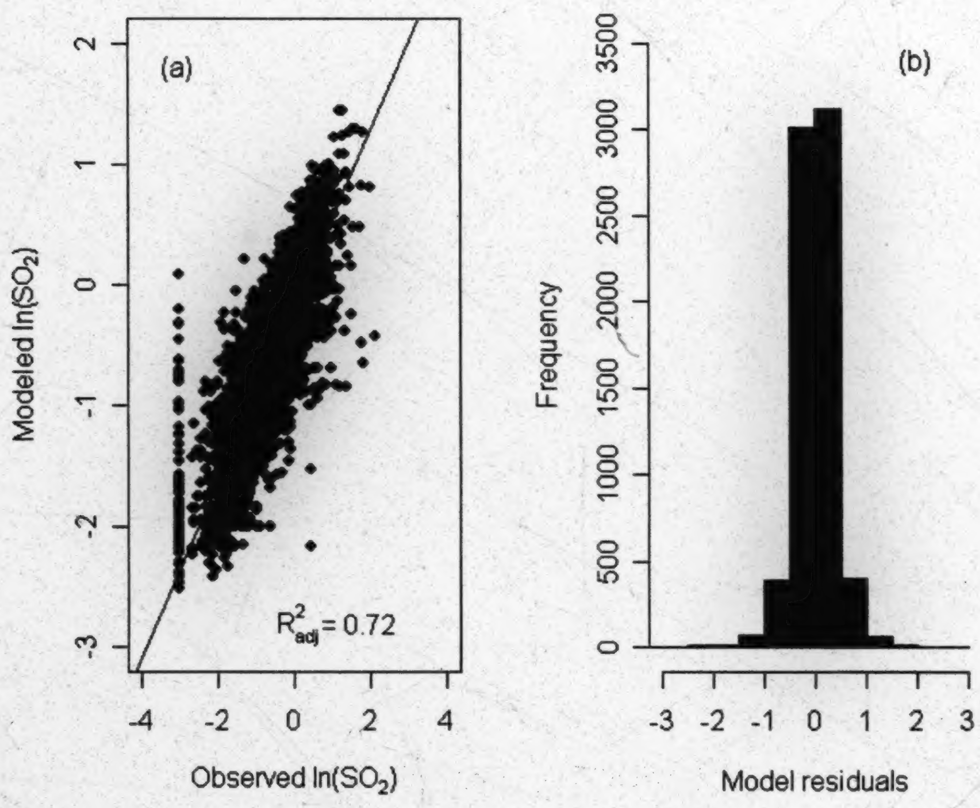
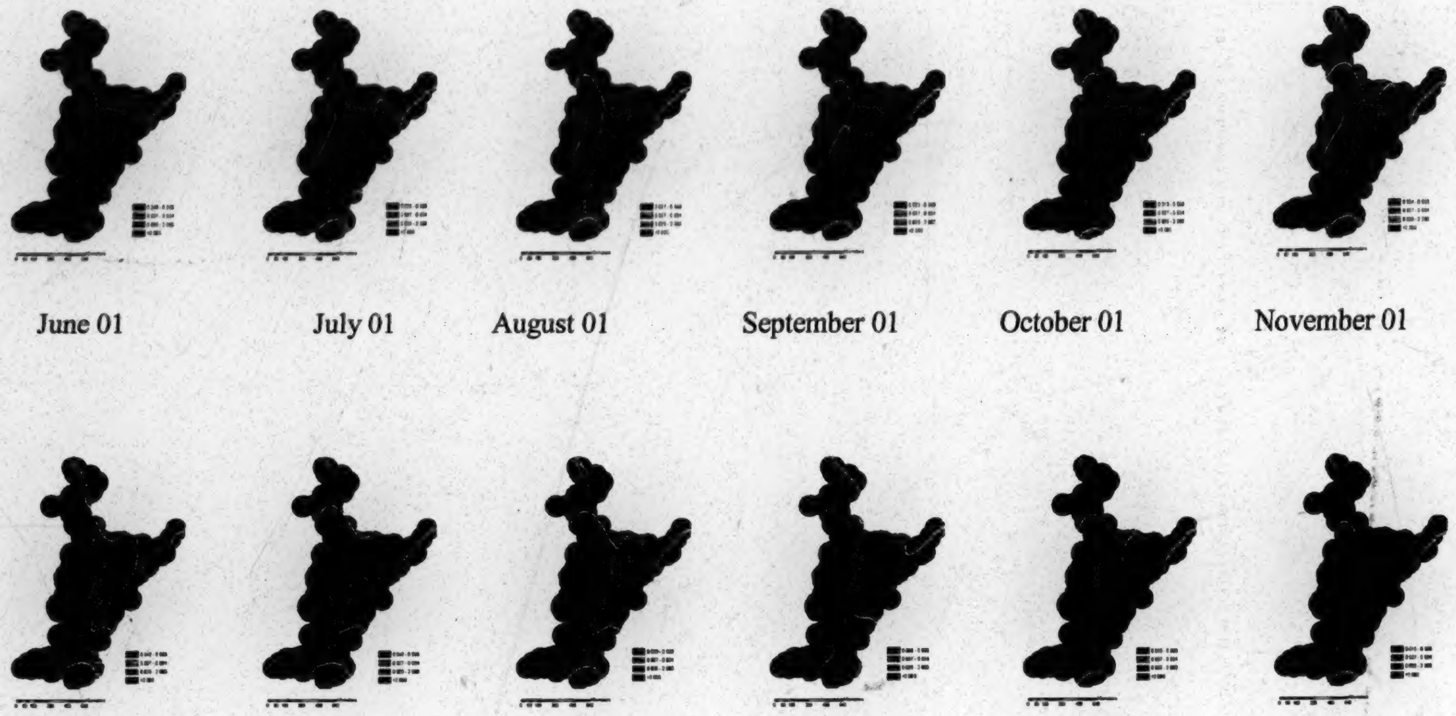


Figure 3.8. Smoothed mean surface maps of ambient sulfur dioxide concentrations by month from Jane 2001 to May 2002.



June 01

July 01

August 01

September 01

October 01

November 01

Figure 3.9. Two hundred simulated Thiessen polygons (irregular polygons in black) and 2000 simulated random points (green dots) overlapping with each of the polygons for prediction of sulfur dioxide exposure in each polygon.

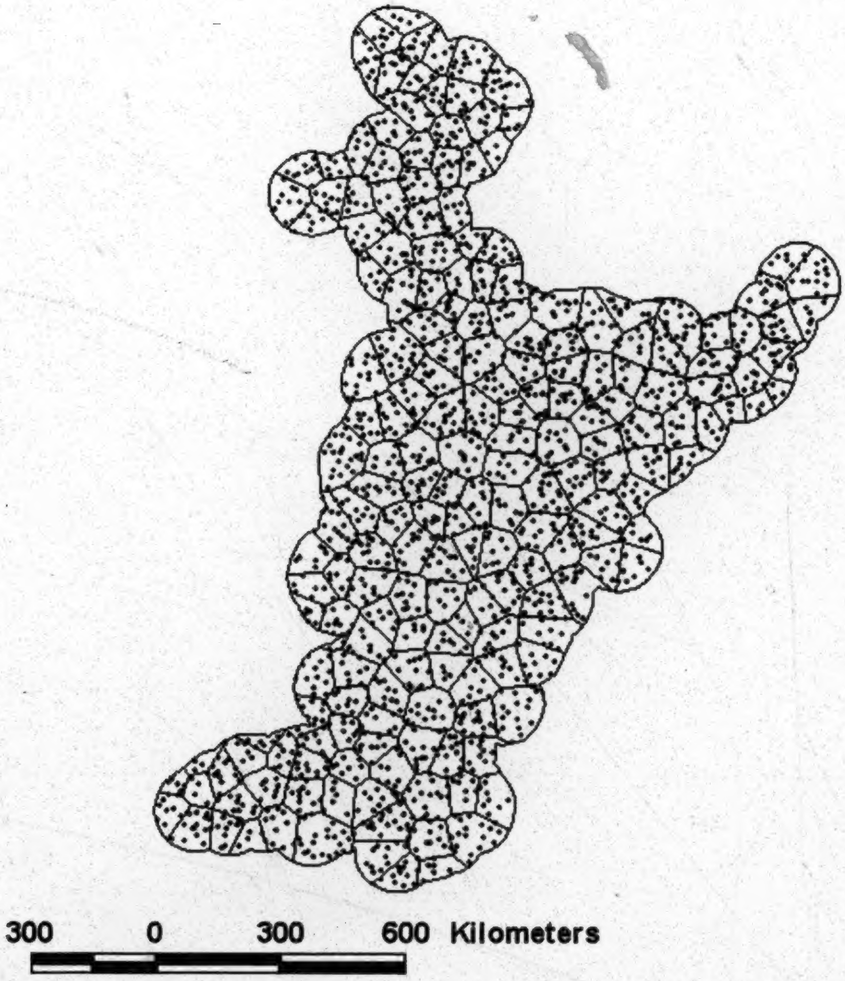


Figure 3.10. Relationship between polygon area and number of points for the 200 simulated polygons.

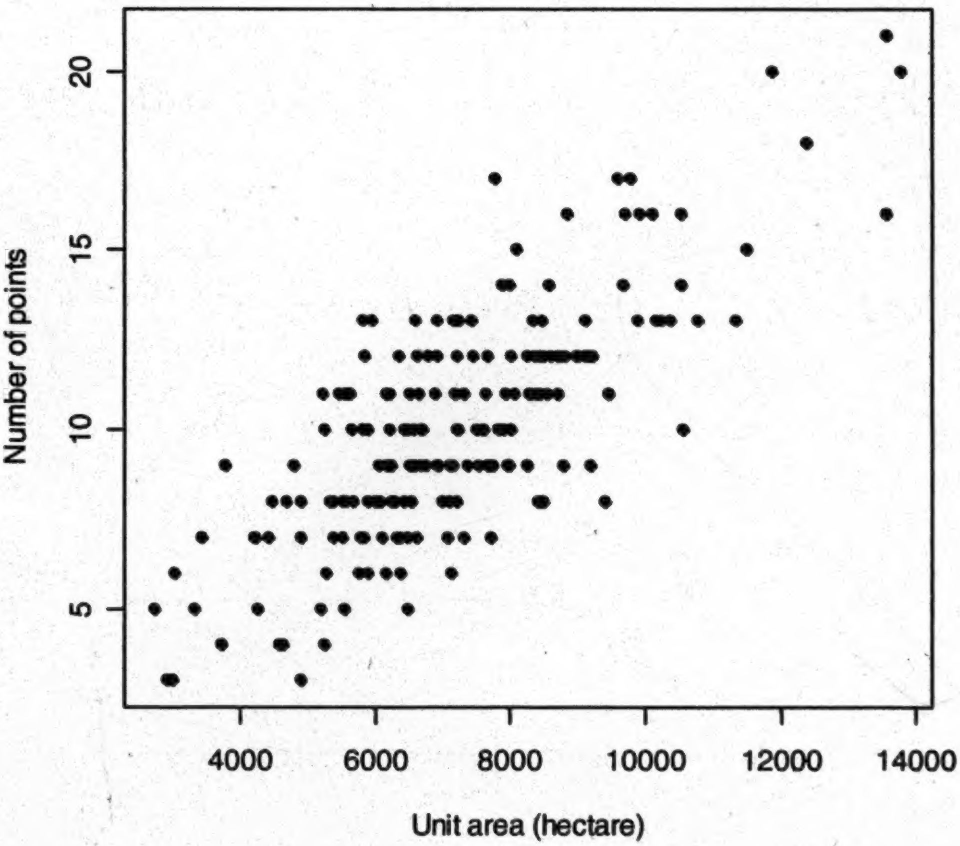


Figure 3.11. Predicted ambient sulfur dioxide concentrations (in ppb) from the spatiotemporal model that were averaged over time and space for each of the 200 polygons.

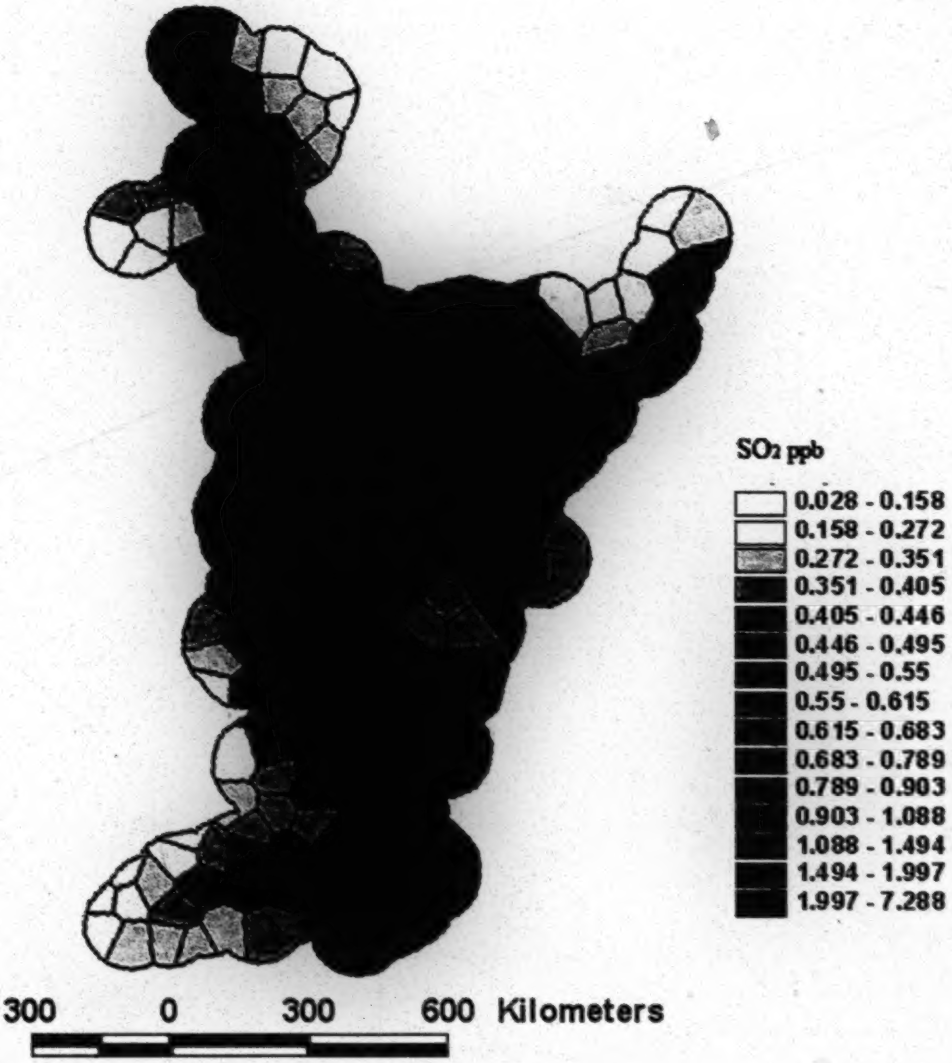

SO2 ppb
0.028 - 0.158
0.158 - 0.272
0.272 - 0.351
0.351 - 0.405
0.405 - 0.446
0.446 - 0.495
0.495 - 0.55
0.55 - 0.615
0.615 - 0.683
0.683 - 0.789
0.789 - 0.903
0.903 - 1.088
1.088 - 1.494
1.494 - 1.997
1.997 - 7.288
300
0
300
600
Kilometers

Figure 3.12. Hierarchical clusters of the 200 polygons, which were formed according to the sulfur dioxide concentrations predicted from the spatiotemporal model for each of the 12 months. Numbers on the X-axis are the sequential polygon numbers. The height on the Y-axis is the value of the criterion for the particular cluster agglomeration.

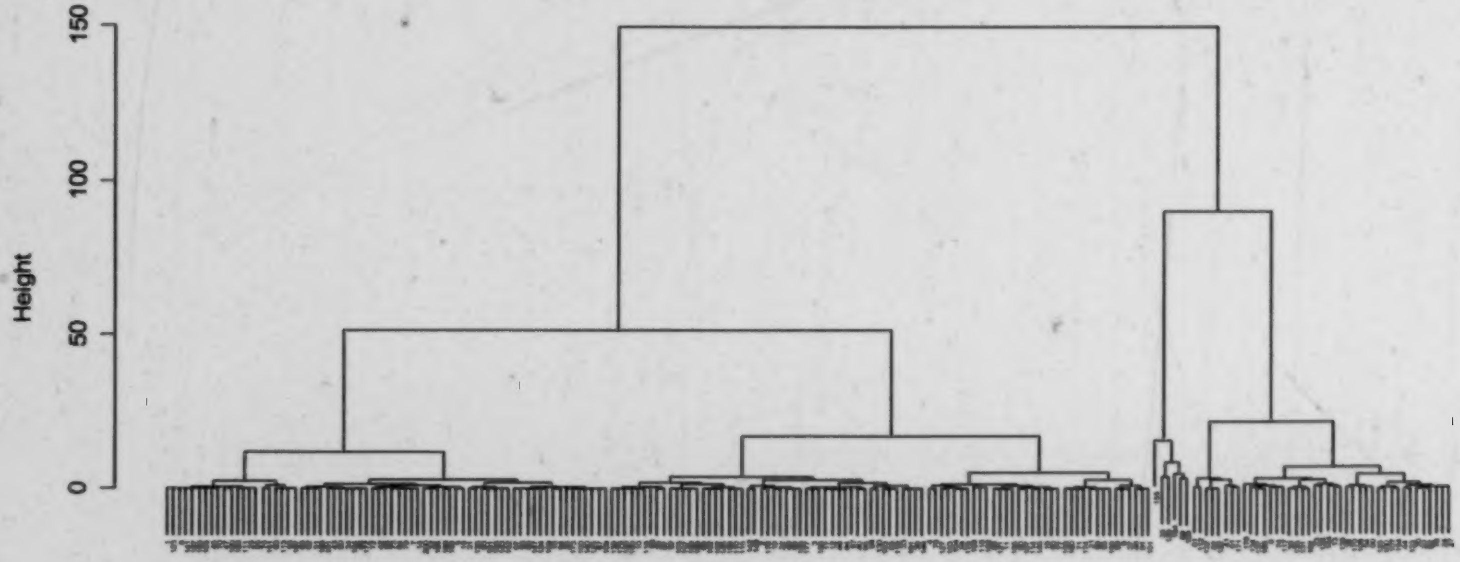


Figure 3.13. Clustered classes of the original 200 polygons with different means of predicted ambient sulfur dioxide concentrations.

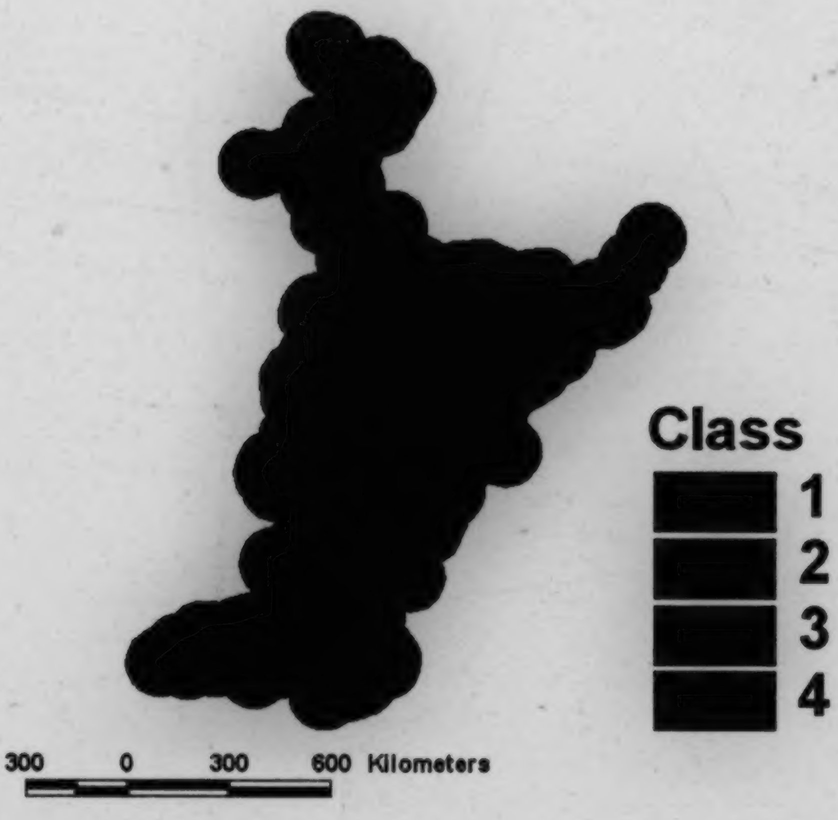
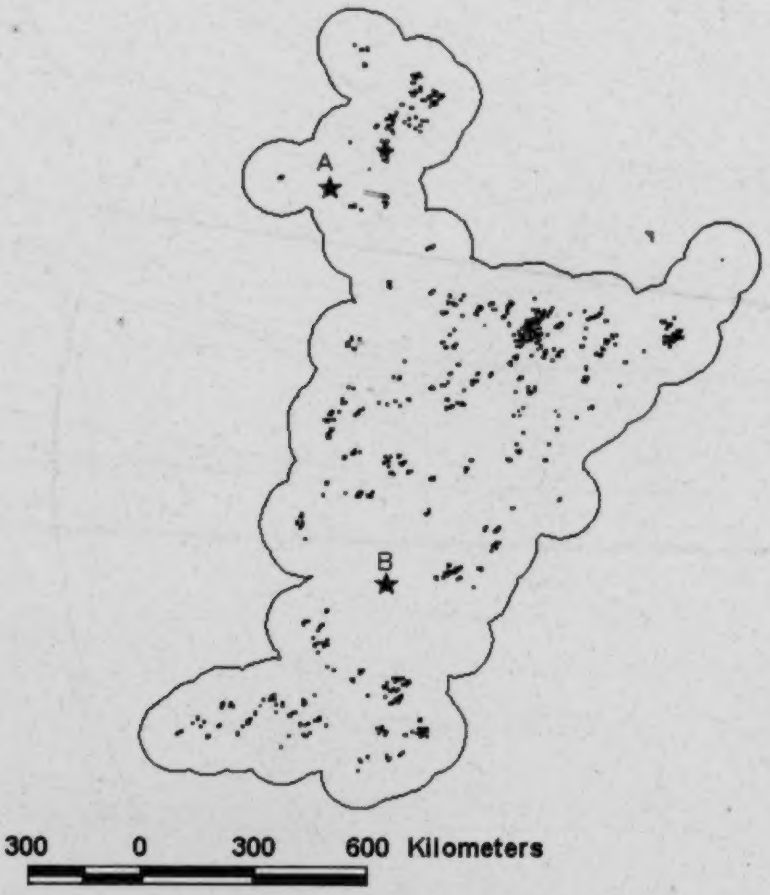

Class
1
2
3
4
300
0
300
600
Kilometers

Figure 3.14. Two random locations (A and B) whose ambient sulfur dioxide concentrations are to be predicted from the spatiotemporal model.


A
B
300
0
300
600 Kilometers

Chapter 4

Discussion and future research

Western Canada is an oil and gas production region. Thermal power generation is also an important industry in both Alberta and Saskatchewan. In these industries, SO_2 and other noxious gases are emitted to the atmosphere in Canada (Environment Canada 2002, Alberta Environment 2005a). An understanding of the spatial, temporal, and spatiotemporal patterns of the ambient SO_2 concentrations is very useful in conducting epidemiological studies on the health effect of outdoor air pollution in Western Canada, despite the apparently low ambient concentrations of SO_2 in this region (Table 3.1, Figure 3.8) as compared to the national standard (Table 2.1). In this thesis, a spatiotemporal model was proposed to predict the ambient SO_2 concentrations and describe spatial and temporal patterns in Western Canada.

The model thus developed is versatile. It can be used to decipher the spatiotemporal patterns of ambient SO_2 concentrations in the study region. A major advantage of using the modeled information is that one can assign discriminatory SO_2 exposure to the at-risk population in different sub-regions, in comparison to the usual practice of using averaged SO_2 concentrations from centrally located stationary monitoring sites in environmental epidemiological studies (Suh 2003, see examples in Timonen and Pekkanen 1997, Kan and Chen 2003, Lee et al. 2003, Liu et al. 2003, 20004). Exposure averaged over a large study region could contain substantial measurement error with respect to personal exposure, as it does not take into account the spatial and temporal variability of the air pollutants and the relationship of concentrations between the average ambient environment and the specific personal breathing zone (Zeger et al. 2000).

Nevertheless, caution is warranted in using the results from the spatiotemporal model in epidemiological studies of the health effects of exposure to SO_2. Specifically, this has to do with the exposure assessment, the very core of this research. At issue is the true and proxy exposure of the subjects (Rappaport and Kupper 2004). The modeled SO_2

concentrations are for the area of the residence, but not for the actual exposure of any individual person in the residential area. They are therefore an ecological measure -- a proxy to the true individual exposure. To what extend that this proxy is close to the true exposure cannot be ascertained by the modeled results themselves. An independent validation with high quality data is required, so that proper variance structure of the exposure measurements can be assessed (Rappaport and Kupper 2004) and adjustments be made (Gustafson 2004). To this end, the modeled results are more restrictive in comparison to other group-based exposure assessment schemes, because in those cases, one can at least work with some measurements of exposure on some individuals within a group in a particular environmental setting. An interesting finding is that group-based exposure assessment errors tend to attenuate the association of exposure and health outcomes to a lesser extent then more error-prone individual-based exposure assessment when measurement error is additive and group means can be precisely estimated (Armstrong 1998, 2003, Kim et al. 2006).

There are other computer models available, which have been used in Alberta to model short range ambient SO_2 concentrations from "industry accidents", such as sour gas well blowout and natural gas pipeline rupture (Alp et al. 1990), and long range distribution and deposition of SO_2 in Canada, including Western Canada (Environment Canada 1997). Both types of those models have more sophisticated inputs than the spatiotemporal statistical model proposed in this thesis. The first model proposed for the short range distribution of ambient SO_2 concentrations is a modular model that consists of modules of emission rate, jet expression, plume rise, transient release ratio, and passive dispersion as the model inputs. The second model proposed for the long range distribution and deposition of SO_2 comprises two component models, one using a backward trajectory model with meteorological factors as the inputs and the other using a forward "moving-box" chemical model with emissions and deposition as the inputs. Because both models have inputs and outputs coupled, they are capable of on-line modeling and forecasting, if the inputs required by the models are fed to the models on-line. Unfortunately, most such inputs are beyond the reach of epidemiologists and therefore may not be practical in epidemiological studies. Attempts have been made to use this approach in epidemiological studies (Scott et al. 2003), and methodologies of how to use environmental data for health effect investigations have been proposed (Best et al. 2000, Richardson and Best 2003).

In their studies of the effects of sour-gas emissions on cattle health, Scott and colleagues (2003) used two models, one being a Gaussian–dispersion model and the other being a distance-decay model. Both models were centered on the point sources of pollution and the SO_2 emissions from the point sources were estimated from the models. The exposure to SO_2 at a location was then summed from all the sources as an indicator of exposure in epidemiological studies. Because retrospective meteorological data and emission data were used as model inputs, many assumptions, such as constant emission rates and stable meteorological conditions, had to be made, which may not be valid. Furthermore, they modeled only the sour-gas plant emissions, but atmospheric SO_2 is known of multiple origins (Lévêque 2003). Consequently, the predicted SO_2 exposure for a location could be off from the truth.

In contrast, the spatiotemporal model presented here is driven primarily by the SO_2 monitoring data, which are cumulative measures from all sources. Unlike the other models mentioned above, this statistical model does not have any covariates either to model the spatial trend or the temporal evolution. Therefore, it does not rely on the assumptions of peripherals, such as constant emission rates and stable meteorological conditions. Like any other statistical models, it draws conclusions according to actual monitoring data that cover the study region. To what extent the modeled results resemble the truth is therefore a matter of data quality, but nothing else.

The spatiotemporal model proposed in this thesis does have limited capability of forecasting, but the forecast would be based only on the past SO_2 data, not on the current environmental or industrial conditions in the study region. Should there be any significant departure of either the meteorological or the industrial conditions or both from those under which the past SO_2 data were generated, the forecast made from this spatiotemporal model would not be correct. The fact is that this spatiotemporal model lacks covariates so that it can not adjust itself for those condition changes. In comparison, the other models would be able to handle the changes, because they are an intrinsic part of the models.

This brings up another note of caution in using the spatiotemporal model proposed in this thesis. It can not be extrapolated too far from the study period. Extrapolation both forward and backward in time is not reliable, as the confidence intervals tend to increase with time lags in time series models (West and Harrison 1997, Shumway and Stoffer 2000). Spatial extrapolation is strongly discouraged because of the lack of any data support outside of the study region.

These limitations, however, do not diminish the merits of this spatiotemporal model and the modeling approach that was undertaken in this research. The process convolution approach adopted in this thesis has several advantages over other spatiotemporal statistical modeling approaches. The first such advantage is the modeling of non-stationary spatiotemporal data. Because the modeling takes a spatial moving-average approach, the data needs not to be *stationary*. In comparison, some geostatistical models require that the data be spatially stationary (Cressie 1994). Otherwise, pre-treatments, such as trend removal, have to be applied before the data are modeled. The second advantage is the handling of misalignment in spatiotemporal data, which often requires imputation of missing values (Little and Rubin 2002) or complicated mathematic treatment in other modeling approaches (Banerjee et al. 2004). By using separate kernels for data measured in displaced space at each time point, the misalignment problem is easily handled (Calder 2005). The third advantage, which has not been explored in this thesis research, is the ability to accommodate spatial anisotropy (Higdon et al. 1999, Higdon 2002). By anisotropy, we mean the spatial process showing different variance in different directions. This often occurs in geology, in which the geological phenomenon being studied is dictated by the orientation of the rock veins. This can also occur in environmental monitoring data when meteorological conditions distort the spatial distribution of the air pollutants. For instance, a prevailing wind in one direction could result in the air pollutants spatially following the downwind direction. Judging according to how the SO_2 data were collected, it is difficult to speculate how meteorological conditions might have caused anisotropy in this study. The fourth advantage is the handling of large datasets, i.e. the large sample (N) problem in spatial statistics (Wikle and Cressie, 1999, Banerjee et al. 2004). This is probably the most significant advantage of this modeling approach. By reducing the original dimension of N to a new dimension of M for $M << N$, the computation suddenly becomes much easier. Another advantage of the approach is that the random process *X* of the model can be specified as correlated, if a priori knowledge or the data appear to suggest so (Lee et al. 2005).

As with any other statistical modeling, some critical assumptions were made in the development of the spatiotemporal model in this research. These were: a) the SO_2 random process was isotropic; b) the underlying random process *X* at each supporting sites was independent; c) the random process *X* temporally evolved in a random walking fashion. Due to these assumptions, it is conceivable that the current work could be subjected to criticism. Further work is required to verify the validity of these assumptions.

As a result, the spatiotemporal model proposed in this thesis should be viewed as exploratory and the model developed as preliminary. Much could be done to improve the modeling accuracy, i.e. the closeness of the modeled results to the truth of the underlying process. Some areas that could be further explored are the incorporation of anisotropy, the assumption of a correlated random process, a stationary temporal evolution of the random process (i.e. not to use the random walk approach, because a time series of random walk is not stationary), incorporation of other covariates, and the re-programming the estimation procedure to achieve faster convergence of the model parameters.

References

Aikma H. 1978. A method of bivariate interpolation and smooth surface fitting for irregularly distributed data points. ACM Transactions on Mathematical Software, 4:148-164.

Alberta Environment. 2005a. Air – What is monitored. http://www3.gov.ab.ca/env/air/AmbientAirMonitoring/whatismonitored.html#Sulfer Dioxide. Last accessed May 16, 2006.

Alberta Environment. 2005b. State of the environment – Air indicators. http://www3.gov.ab.ca/env/soe/air_indicators/7_sulfurdioxide.html. Last accessed May 16, 2006.

Alp E, Davies MJE, Huget RG, Lam LH, Zelensky MJ. 1990. A model to estimate ground-level H_2S and SO_2 concentrations and consequences from uncontrolled sour gas releases. Concord Environmental Corporation, Calgary, Alberta. Canada, 287 p.

Alvo M, Dabrowski AR. 2000. Measuring regional trends in ozone. Environ Model Assess, 5:217-228.

Armstrong BK, White E, Saracci R. 1993. Principles of exposure measurement in epidemiology. Oxford University Press. New York, NY. 351 p.

Armstrong BG. 1998. Effect of measurement error on epidemiological studies of environmental and occupational exposures. Occup Environ. Med, 55:651-656.

Armstrong BG. 2003. Exposure measurement error: consequences and design issues. In In Exposure assessment in occupational and environmental epidemiology (Nieuwenhuijsen MJ, ed.). Oxford University Press. New York, NY. Pp181-200.

Asgari MM, DuBois A, Asgari M, Gent J, Becket WS. 1998. Association of ambient air quality with children's lung function in urban and rural Iran. Arch Environ Health, 53:222-230.

Baddeley A, Turner R. 2005. Spatstat: an R package for analyzing spatial point patterns. J Stat Software, 12:1-42.

Banerjee S, Carlin BP, Gelfand AE. 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman and Hall/CRC, Boca Raton, Fla. 452 p.

Barnett AG, Williams GM, Schwartz J, Neller AH, Best TL, Petroeschevsky AL, Simpson RW. 2005. Air pollution and child respiratory health: A case-crossover study in Australia and New Zealand. Amer Respir Crit Care Med, 171:1272 – 1278.

Bates D, Maechler M. 2005. Matrix: A Matrix package for R. R package version 0.98-7

Bernstein JA (ed.) 2004. Health effects of air pollution. J. Allergy Clin Immunol, 114: 1116-1123.

Bennett RJ. 1979. Spatial time series: Analysis-forecasting-control. Pion, London, UK, 674 p.

Besag J. 1974. Spatial interaction and the statistical analysis of lattice systems (with Discussion). J. Royal Stat Soc, Series B, 36:192-236.

Best NG, Ickstadt K, Wolpert RL. 2000. Spatial Poisson regression for health and exposure data measured at disparate resolutions. J Amer Stat Assoc. 95:1076-1088.

Biggeri A, Baccini M, Bellini P, Terracini B. 2005. Meta-analysis of the Italian Studies of Short-term Effects of Air Pollution (MISA), 1990–1999. Int J Occup Environ Health, 11:107–122.

Bocci C, Dabrowski AR. 2002. Fitting data to linear combinations of covariograms. Environ Model Assess, 7: 39-46.

Brook JR, Maker PA, Moran MD, Shepherd MF, Vet RJ, Dann TF, Dion J. 2001. Precursor contributions to ambient fine particulate matter in Canada. Environment Canada. http://www.msc-smc.ec.gc.ca/saib/smog/pm _full/ pm2_5_full_pg14 _e.html#4.4. Last accessed May 20, 2006.

Brook RD, Brook JR, Rajagopalan S. 2003. Air pollution: the “heart” of the problem. Curr Hypertens Rep, 5:32–39.

Brook RD, Franklin B, Cascio W, Hong Y, Howard G, Lipsett M, Luepker R, Mittleman M, Samet J, Smith, SC Jr, Tager I. 2004. Air Pollution and Cardiovascular Disease: A Statement for Healthcare Professionals From the Expert Panel on Population and Prevention Science of the American Heart Association. Circulation, 109:2655-2671.

Brunekreef B, Holgate ST. 2002. Air pollution and health. Lancet, 360: 1233–1242.

Brown PE, Karesen KF, Roberts GO, Tonellato S. 2000. Blur-generated non-separable space-time models. J Royal Stat Soc, Series B, 62:847-860.

Brunce N. 1994. Environmental Chemistry, 2nd ed. Wuerz Publishing Ltd., Winnipeg, Manitoba, 376 p.

Burstyn I, Senthilselvan A, Cherry N. 2006. Statistical analysis of the anthropogenic determinants of concentrations of air pollutants. *In* Western Canada study of animal health effects associated with exposure to emissions from oil and natural gas field facilities: Research appendices. Western Interprovincial Scientific Studies Association. Pp. 4.1-4.41.

Calder CA, 2003. Exploring Latent Structure in spatial temporal process using process convolutions. Institute of Statistics and Decision Sciences, Duke University. Ph.D. Thesis.

Calder CA, 2005. Dynamic Factor Process Convolution Models for Multivariate Space-Time Data with Application to Air Quality Assessment. Department of Statistics Preprint No. 750, The Ohio State University.

Calder CA, Lavine M, Müller P, Clark JS. 2003. Incorporating multiple sources of stochasticity into dynamic population models. Ecology, 84:1395-1402.

Calder CA, Holloman C, and Higdon D. 2001. Exploring space-time structure in ozone concentration using a dynamic process convolution model. Institute of Statistics & Decision Sciences, Duke University, Durham, USA.

Calvert JG, Stockwell WR. 1984. Mechanisms and rates of gas-phase oxidations of sulfur dioxide and nitrogen oxides in the atmosphere. *In* SO2, NOx, oxidation mechanisms: Atmospheric considerations (Calvert JG (ed.)). Butterworths, Toronto, ON, Pp. 1-62.

Carroll R, Chen R, Li T, Newton H. Schmiediche H, Wang H, George E. 1997. Modeling ozone exposure in Harris County, Texas. J. Am Statist Assoc, 92:392-413.

Carter CK, Kohn R. 1994. On Gibbs Sampling for State Space Models. Biometrika, 81:541-553.

Chen M, Shao Q, Ibrahim JG. 2000. Monte Carlo methods in Bayesian computation. Springer-Verlag, New York, NY, 386 p.

Chew FT, Goh DY, Ooi BC, Saharom R, Hui JK, Lee BW. 1999. Association of ambient air-pollution levels with acute asthma exacerbation among children in Singapore. Allergy, 54:320-329.

Cressie NAC. 1993. Statistics for Spatial data. John Wiley & Sons, New York, NY. 928 p.

Cressie NAC, Wikle CK. 2002. Space-time Kalman Filter. *In* Encyclopedia of Environmetrics (El-Shaarawi AH, Piegorsch WW (eds.)). Vol. 4, Pp 2045–2049. John Wiley & Sons, Chichester.

Dales R, Burnnet RT, Smith-Doiron M, Stieb DM, Brook JR. 2004. Air pollution and sudden infant death syndrome. Pediatrics, 113:628-631.

Davies M, Prasad S, Gieni W, Vanderheydan M, Vatcher C, Whiteley S, Bhardwa S. 2006. Air-concentration monitoring technology, methods and overview. In Western Canada study of animal health effects associated with exposure to emissions from oil and natural gas field facilities: Research appendices. The Western Interprovincial Scientific Studies Association. Pp 3.1-3.61.

De Iaco S, Myers DE, Posa D. 2002. Space–time variograms and a functional form for total air pollution measurements. Comput Stat Data Anal, 41:311 – 328.

De Iaco S, Myers DE, Posa D. 2003. The linear coregionalization model and the product-sum space-time variogram. Math Geol, 35:25-37.

Echer MD, Gelfand AE. 1999. Bayesian modeling and inference for geometrically anisotropic spatial data. Mathematical Geology, 31:67-83.

Elliott P, Wakefield J, Best N, Briggs D (eds.). 2000. Spatial epidemiology: methods and applications. Oxford University Press, New York. 494 p.

Environment Canada. 1997. The 1997 Canada Acid Rain Assessment, Vol. 2. Atmospheric Service Assessment Report. Meteorological Services Canada. Environment Canada, Downsview, Ontario. 296 p.

Environment Canada 2002. 2001 annual progress report on the Canada-Wide Acid Rain Strategy for Post-2000. (http://www.ccme.ca/assets/pdf/acid_rain_e.pdf)

Environmental Criteria and Assessment Office. 1982. Air Quality Criteria for Particulate matter and sulfur oxides, Vol. 1. U.S. Environmental Protection Agency, Office of Research and Development, Office of Health and Environment Assessment. Washington, D.C.

Environmental Systems Research Institute, Inc. (ESRI). 1999. ArcView 3.2. Redlands, CA.

ESRI, Inc. 2005. ArcGIS 9.1.

Everitt B, Rabe-Hesketh S. 2001. Analyzing medical data using S-plus. Spinger, New York, New York. Pp 243-290.

Farwell S, Chatham WH, Barinaga CJ. 1987. Performance characterization and optimization of the $AgNO_3$-filter/FMA Fluorometric method for atmospheric H_2S. measurements. J Air Pollution Control Assoc, 37:1052-1059.

Federal-Provincial Advisory Committee on Air Quality. 1987. Review of national ambient air quality objectives for sulfur dioxide: Desirable and acceptable levels. Ministry of Supply and Services Canada. 36 p. Beauregard Press Ltd.

Fernéndez-Casal R, González-Manteiga W, Febrero-Bande. 2003. Space-time dependency modeling using general classes of flexible stationary variogram models. J Geophys Res, 108(D24), 8779, doi:10.1029/2002JD002909.

Ferris BJ Jr., Ware JH, Speizer FE. 1980. Review of SOx and particulate standard: The epidemiologic evidence. In Proceedings: The proposed SO_x and particulate standard specialty conference (The Air Pollution Control Association, ed.). Southern section, Air Pollution Control Association, September 1980.Atlanta, Georgia. Pp 240 – 247.

Firket J. 1936. Fog along the Meuse Valley. Trans Faraday Soc, 32:1102-1197 (cited in National Air Pollution Control Administration 1969).

Frank R. 1980. The toxicity of sulfur oxides. In Proceedings: The proposed SOx and particulate standard specialty conference (The Air Pollution Control Association, ed.). Southern section, Air Pollution Control Association, September 1980.Atlanta, Georgia. Pp 183 – 203.

Frank NR, Speizer F. 1964 Uptake and release of SO2 by the human nose. Physiol, 7:132.

Friend JP. 1973. The global sulfur cycle. *In* Chemistry of the lower troposphere (Rasool SI (ed.)). Plenum Press, New York, NY, Pp 177-201.

Frühwirth-Schnatter S. 1994. Data augmentation and dynamic linear models. J Time Series Anal, 15:183-202.

Fuentes M. 2002a. Spectral methods for nonstationary spatial processes. Biometrika, 89: 197-210.

Fuentes M. 2002b. Modeling and prediction of non-stationary spatial processes. Statist Model, 2:281-298.

Gamerman D, Migon HS. 1993. Dynamic hierarchical models. J R Statist Soc, B, 55:629–642.

Gelfand AE, Smith AFM. 1990. Sampling-based approaches to calculating marginal densities. J Amer Stat Assoc, 85:129-133.

Gelman A, Rubin D. 1992. Inference from iterative simulation using multiple sequences. Statist Sci, 7:457–511.

Gelman S, Gelman D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern analysis and Machine Intelligence, 6:721-741.

Gelman A, Carlin JB, Stern HS, Rubin DB. 1995. Bayesian Data Analysis. Chapman & Hall. New York, NY. 542 p.

Gilks WR, Richardson S, Spiegelhalter DJ. 1996. Markov Chain Monte Carlo in Practice. Chapman and Hall. New York, NY. 486 p.

Gustafson P. 2004. Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments. Chapman and Hall/CRC. New York, NY. 188 p.

Guttorp P, Meiring W, Sampson P. 1994. A space-time analysis of ground-level ozone data. Environmetrics, 5:241–254.

Handcock M, Wallis J. 1994. An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). J Amer Statist Assoc, 89:368–390.

Harrison RM. 1990. Chemistry of the troposphere. In Pollution: Causes, Effects and Control (2nd ed.) (Harrison RM (ed.)). CRC Press, Coca Raton, FL, Pp 157-180.

Haslett J, Raftery AE. 1989. Space-time modeling with long memory dependence: Assessing Ireland's wind power resources (with discussion). Appl Stat, 38:1-50.

Hedley AJ, Wong C, Thach TQ, Ma S, Lam T, Anderson HR. 2002. Cardiorespiratory and all-cause mortality after restrictions on sulfur content of fuel in Hong Kong: an intervention study. Lancet, 360:1646–1652.

Herbarth O, Fritz G, Krumbiegel P, Diez U, Franck U, Richter M. 2001.Effect of sulfur dioxide and particulate pollutants on bronchitis in children--a risk analysis. Environ Toxic, 16:269-76.

Hidy GM. 1994. Atmospheric sulfur and nitrogen oxides: Eastern North America Source-receptor Relationships. Academic Press, San Diego. 447p.

Higdon D. 1998. A process-convolution approach to modeling temperatures in the north Atlantic Ocean. J Environ Ecol Stat, 5:173-190.

Higdon D. 2002. Space and space-time modeling using process convolutions. *In* Quantitative methods for current environmental issues (Anderson C, Barnett V, Chatwin PC, El-Shaarawi AH (eds.)). Springer Verlag. Pp. 37-56.

Higdon D, Swall J, Kern J. 1999. Nonstationary spatial modeling. *In* Bayesian Statistics 6 (Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds.)). Oxford University Press, Oxford, England. Pp 761-768.

Hornung RW, Reed LD. 1990. Estimation of average concentration in presence of nondetectable values. Appl Occup Environ Hyg, 5:46-51.

Huang H, Cressie N. 1996. Spatio-temporal prediction of snow water equivalent using the Kalman filter. Comput Statist Data Anal, 22:159–175.

Huerta G, Sansó B, Stround JR. 2004. A spatiotemporal model for Mexico City ozone levels. Appl Statist, 53 (Part 2):231–248.

Ibanez F, Grosjean P, Etienne M. 2006. pastecs: Package for Analysis of Space-Time Ecological Series. R package version 1.3-1. http://www.sciviews.org/pastecs.

Kalman RE. 1960. A new approach to linear filtering and prediction problems. J Basic Engineering, 82:34-45.

Kan HD, Jia I, Chen BH. 2004. The association of daily diabetes mortality and outdoor air Pollution in Shanghai, China. J Environ Health, 67:21-25.

Kan HD, Chen BH. 2003. Air Pollution and daily mortality in Shanghai: A time series study. Arch Environ Health, 58:360-367.

Katz M. 1977. The Canadian sulfur problem. *In* Sulfur and its inorganic derivatives in the Canadian Environment. Ad hoc Panel of Experts Management Subcommittee, NRC Associate Committee on Scientific Criteria for Environmental Quality, National Research Council of Canada, Ottawa, ON, Pp 21-67.

Kern, JC. 2000. Bayesian process-convolution approaches to specifying spatial dependence structure. Ph.D. dissertation, Institute of Statistics & Decision Sciences, Duke University, Durham, USA.

Kim H, Yasui Y, Burstyn I. 2006. Attenuation in risk estimates in logistic and Cox proportional-hazards models due to group-based exposure assessment strategy. Ann Occup Hyg. May 2, 2006 advanced publication.

Komarnisky LA, Christopherson RJ, Basu TK. 2003. Sulfur: Its Clinical and Toxicologic Aspects. Nutrition 19: 54-61.

Koren HS. 1995. Associations between criteria air pollutants and asthma. Environ. Health Perspect, 103 (Suppl. 61995):235-242.

Koren H, O'Neill M. 1998. Experimental assessment of the influence of atmospheric pollutants on respiratory disease. Toxic Letters, 103:317-321.

Kyriakidis PC, Journel AG. 1999. Geostatistical space-time models: A review. Math Geol, 31:651-683.

Lavine M, Lozier S. 1999. A Markov random field spatio-temporal analysis of ocean temperature. Environ Ecol Stat, 6:249-273.

Lawther PJ, Waller RE, Henderson M. 1970. Air pollution and exacerbations of bronchitis. Thorax, 25:525-539.

Lee J, Berger JO. 2003. Space–time modeling of vertical ozone profiles. Environmetrics, 14:617–639.

Lea J, Kim H, Hong Y, Ha E, Park H. 2003. Air Pollution and Hospital Admissions for Ischemic Heart Diseases among Individuals 64+ Years of Age Residing in Seoul, Korea. Arch Environ Health, 58:617-623.

Lee HKH, Higdon DM, Calder CA, Holloman CH. 2005. Efficient models for correlated data via convolutions of intrinsic processes. Stat Modeling, 5:53-74.

Lévêque C. 2003. Ecology: From ecosystem to Biosphere. Science Publishers, Inc. Enfield, USA. pp. 342-344.

Lin CA, Pereira1 LAA, Nishioka DC, Conceição GMS, Braga ALF, Saldiva PHN. 2004. Air pollution and neonatal deaths in São Paulo, Brazil. Neonatal deaths and air pollution. Brazil J Med Biol Res, 37:765-770

Lin M, Chen Y, Burnett RT, Villeneuve PJ, Krewski D. 2003. Effect of short-term exposure to gaseous pollution on asthma hospitalization in children: a bi-directional case-crossover analysis. J Epidemiol Community Health, 57:50–55

Little RJA, Rubin DB. 2002. Statistical analysis with missing data (2nd ed.). Hoboken, N.J. John Wiley & Sons. 381 p.

Liu S, Krewski D, Shi Y, Chen Y, Burnett RT. 2003. Association between gaseous ambient air pollutants and adverse pregnancy outcomes in Vancouver, Canada. Environ Health Perspect, 111:1773-1778.

Liu S, Krewski D, Shi Y, Chen Y, Burnett RT. 2004. Air pollution and adverse pregnancy outcomes: Response. Environ Health Perspect, 112:A792-794.

Lubin JH, Colt JC, Camann D, Davis D, Cerhan JR, Severson RK, Bernstein L, Hartge P. 2004. Epidemiologic Evaluation of Measurement Data in the Presence of Detection Limits. Environ Health Perspect, 112:1691-1696.

Mardia KV, Goodall CR. 1993. Spatial-temporal analysis of mulitvariate environmental monitoring data. *In* Multivariate Environmental Statistics (Patil GP, Rao CR (eds.)), Elsevier Science Publishers, Pp 347-386.

Mardia K, Goodall C, Redfern E, Alonso F. 1998. The kriged Kalman filter. Test, 7: 217–276.

Martin AE. 1964. Mortality and morbidity statistics and air pollution. Proc R Soc Med, 57:969-975.

Mateu J, Montes F, Fuentes M. 2003. Recent advances in space-time statistics with applications to environmental data: An overview. J Geophys Res 108 (D24), 8774, doi:101029/2003JD003819.

Matheron G. 1963. Principles of geostatistics. Econ Geol, 58:1246-1266.

McManus MS, Koenig JQ, Altman LC, Pierson WE. 1989. Pulmonary effects of sulfur dioxide exposure and ipratropium bromide pretreatment in adults with nonallergic asthma. J Allergy Clinic Immunol, 3:19-626

McMillan N, Bortnick SM, Irwin ME, Berliner LM. 2005. A hierarchical Bayesian model to estimate and forecast ozone through space and time. Atmos Environ, 39:1373–1382.

Meinhold RJ, Singpurwalla NR. 1983. Understanding the Kalman Filter. Amer Stat, 37:123-127.

Meyn SP, Tweedie RL 1993. Markov Chains and Stochastic Stability. Springer-Verlag. London.

Michaud J, Krupitsky D, Grove JS, Anderson BS. 2005. Volcano related atmospheric toxicants in Hilo and Hawaii Volcanoes National Park: Implications for human health. NeuroToxicology, 26:555–563.

Myer R, Millar RB. 1999. BUGS in Bayesian stock assessments. Can J Fish Aquat Sci, 56:1078-1086.

National Air Pollution Control Administration 1969. Air quality criteria for sulfur dioxides. U.S. Department of Health, Education, and Welfare, Public Health Service, Consumer Protection and Environmental Health Services. Washington, D.C. National Air Pollution Control Administration Publication No. AP-50.

Nowak D, Jörres R, Berger J, Claussen M, Magnussen H. 1997. Airway responsiveness to sulfur dioxide in an adult population sample. Amer J Respir Crit Care Med, 156:1151–1156.

Oleckno WA. 2002. Essential Epidemiology: Principles and applications. Waveland Press, Inc. Long Grove, Ill. 368 p.

Pinheiro J, Bates D, DebRoy S, Sarkar D. 2006 . nlme: Linear and nonlinear mixed effects models. R package version 3.1-68.1.

Plummer M, Best N, Cowles K, Vinescoda K. 2006. coda: Output analysis and diagnostics for MCMC. R package version 0.10-5. http://www-fis.iarc.fr/coda/

Pope CA. 2000. Epidemiology of fine particulate air pollution and human health: biologic mechanisms and who's at risk? Environ Health Perspect, 108:713–723.

R Development Core Team. 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Rappaport SM, Kupper LL. 2004. Variability of environmental exposures to volatile organic compounds. J Exp Anal Environ Epidemiol, 14:92-107.

Ribeiro JR , Diggle PJ. 2001. geoR: A package for geostatistical analysis. R-NEWS Vol 1, No 2.

Riccio A. 2005. A Bayesian approach for the spatiotemporal interpolation of environmental data. Monthly Weather Review, 133:430-440.

Richardson S, Best N. 2003. Bayesian hierarchical models in ecological studies of health–environment effects. Environmetrics, 14:129-147.

Rodriguez-Iturbe I, Mejia JM. 1974, The design of rainfall networks in time and space. Water Resources Res, 10:713-729.

Romesburg HC. 1984. Cluster analysis for researchers. Lifetime Learning Publications, Belmont, Calif. Pp119 – 140.

Routledge HC, Ayres JG. 2005. Air pollution and the heart. Occup Med, 55:439–447.

Sahu SK, Mardia KV. 2005. A Bayesian kriged Kalman model for short-term forecasting of air pollution levels. Appl Statist, 54 (*Part* 1): 223–244

Sampson P, Guttorp P. 1992. Nonparametric estimation of nonstationary spatial covariance structure. J Amer Statist Assoc, 87:108–119.

Sansó B, Guenni L. 2000. A nonstationary multisite model for rainfall. J Amer Statist Assoc, 95:1089 – 1100.

Srám RJ, Binková B, Dejmek J, Bobak M. 2005. Ambient Air Pollution and Pregnancy Outcomes: A Review of the Literature. Environ Health Perspest, 113:375-382

Schrenk HH, Heimann H, Clayton GO, Gafafer WM, Wexler H. 1949. Air pollution in Donora, Pennsylvania, Epidemiology of the unusual smog episode of October 1948. Pub Health Bull No. 306. Fed Sec.Agency, Washington, D.C. (cited in National Air Pollution Control Administration 1969).

Schwartz J. 1993. Air pollution and daily mortality in Birmingham, Alabama. Amer J Epidem, 137:1136–1147.

Schwartz J, Dockery D. 1992. Increased mortality in Philadelphia associated with daily air pollution concentrations. Amer Rev Respir Dis, 145:600–604.

Scott, HM, Soskolne CL, Martin SW, Ellehoj EA, Coppock RW, Guidotti TL, Lissemore TL. 2003. Comparison of two atmospheric-dispersion models to assess farm-site exposure to sour-gas processing plant emissions. Prev Vet Med, 57:15-34.

Sembulak S, Kindzierski W. 1999. Rural and urban passive monitoring of sulfur dioxide. Prepared for the Sustainable Forest Management Network Project: Exposure

Assessment of Air Pollutants from the Forest Industry by Department of Civil Engineering, University of Alberta. Project Report 1999-2. 19 p.

Shaddick G, Wakefield J. 2002. Modeling daily multivariate pollutant data at multiple sites. Appl Statist, 51:351-372.

Shumway RH, Stoffer DS. 2000. Time series analysis and its applications. Springer-Verlag, New York, NY. 549 p.

Silverman B. 1986. Density estimation for statistics and data analysis. Boca Raton, FL. Chapman and Hall/CRC Press.

Smith AFM, GO R. 1993. Bayesian computation via the Gibbs sampler and other related Markov chain Monte Carlo methods. J R Statist Soc B, 55:3–23.

Smith RL, Kolenikov S, Cox LH. 2003. Spatiotemporal modeling of PM2.5 data with missing values. J Geophys Res, Vol. 108, No. D24, 9004, doi:10.1029/2002JD002914.

Speizer F, Frank NR. 1966. The uptake and release of SO_2 by the human nose. Arch Environ Health, 12:725-728.

Spiegelhalter D, Thomas A, Best N, Lunn D. 2003, WinBUGS 1.4 Manual.

Spiegelhalter D, Thomas A, Best N. 1996. Computation on Bayesian graphical models. *In* Bayesian Statistics 5 (Bernardo JM, Berger JO, Dawid AP and Smith AFM (eds.)). Oxford University Press, Oxford, UK. Pp 407-425.

Stacy RW, Friedman M, Hazucha M, Green J. 1981. Effects of 0.75 ppm sulfur dioxide on pulmonary function parameters of normal human subjects. Arch Environ Health, 36:172-178.

Suh HH. 2003. Particulate matter. In Exposure assessment in occupational and environmental epidemiology (Nieuwenhuijsen MJ, ed.). Oxford University Press. New York, NY. Pp 221-236.

Tang H, Brassard B, Brassard R, Peake E.1997. A new passive sampling system for monitoring SO_2 in the atmosphere. Field Anal Chem Tech, 1:307-314.

Tielemans E, Kupper L, Kromhout H, Heederik D, Houba R. 1998. Individual-based and group-based occupational exposure assessment: some equations to evaluate different strategies. Ann Occup Hyg, 42:115-119.

Timm NH. 2002. Applied multivariate analysis. Springer, New York, NY. Pp 515-556.

Timonen KL, Pekkanen. 1997. Air Pollution and Respiratory Health among Children with Asthmatic or Cough Symptoms. Amer J Respir Crit Care Med, 156:546–552.

Tonellato SF. 2001. A multivariate time series model for the analysis and prediction of carbon monoxide atmospheric concentrations. Appl Statist, 50:187–200.

United Nations, Economic Commission for Europe, 1985. Air pollution across boundaries: Report prepared within the framework of the Convention on Long-range Transboundary Air Pollution. United Nations Publication Sale No. E85.II.E17.

Von Burg R. 1995. Toxicological update. J Appl Toxicol, 16:365-371.

Waldbott GL. 1978. Health effects of environmental pollutants. The C.V. Mosby Company, St. Louis, Missouri. Pp 86-89.

Waller LA, Gotway CA. 2004. Applied spatial statistics for public health data. Wiley-Interscience. Hoboken, NJ. Pp.272-324.

WBK & Associates Inc., 2003. Sulfur dioxide: Environmental effects, fate and behavior. Alberta Environment, Edmonton, Alberta. http://www.gov.ab.ca/env.

West M, Harrison J. 1997. Bayesian forecasting and dynamic models (2nd ed.). New York, New York, Springer-Verlag. 568 p.

Western Research & Development. 1978. Derivation of first order estimates of sulfur deposition in the region of representative point sources. Alberta Environment, Edmonton, Alberta. Ref: 2737-R1.

Wikle C, Cressie N. 1999. A dimension reduction approach to space-time Kalman filtering. Biometrika, 86:815–829.

Wikle C, Berliner L, Cressie N. 1998. Hierarchical Bayesian space-time models. Environ Ecol Statist, 5:117–154.

Wilkins ET. 1954. Air pollution aspects of the London fog of December 1953. Royal Meteorol Soc J, 80: 267-271 (cited in National Air Pollution Control Administration 1969).

Wilson AM, Cameron P, Wake CP, Kelly T, Salloway JC. 2005. Air pollution, weather, and respiratory emergency room visits in two northern New England cities: an ecological time-series study. Environ Res, 97:312–321.

Wolf RK, Dolovich M, Rossman CM, Newhouse MT. 1975. Sulfur dioxide and tracheobronchial clearance in man. Arch Environ Health, 30:521-527.

Wong TW, Tam WS, Yu TS, Wong AH. 2002. Associations between daily mortalities from respiratory and cardiovascular diseases and air pollution in Hong Kong, China. Occup. Environ Med, 59:30-5.

Wong GW, Ko FW, Lau TS, Li ST, Hui D, Pang SW, Leung R, Fok TF, Lai CK. 2001.Temporal relationship between air pollution and hospital admissions for asthmatic children in Hong Kong. Clin Expt Allergy, 31:565-569.

Xu X, Li B, Huang H. 1995. Air pollution and unscheduled hospital outpatient and emergency room visits. Environ Health Perspect, 103:286-289.

Zeger SL, Thomas D, Dominici F, Samet JM, Schwartz J, Dockery D, Cohen A. 2000. Exposure measurement error in time-series studies of air pollution: concepts and consequences. Environmental Health Perspective, 108:419-426.

Appendix I

Glossary

Areal data: Also called lattice data. They are a type of spatial data that are aggregated based on areal units, such as a census tract, a postal code coverage, an electoral riding, a county, a province. In some applications, continuous data may be converted into areal data by dividing the study region into equal sized cells through a grid system.

Markov Chain: A mathematical concept. On a countable state space, a discrete time stochastic process whose conditional probability distribution in the next future state (t = t+1), give the present and past states, is dependent only on the present (t = t) state is a Markov chain. This definition on the discrete space can be extended to a more general space.

Markov Chain Monte Carlo: Mathematical methods (algorithms) that are based on constructing a Markov Chain of certain desired stationary properties for repeatedly sampling to arrive a sample that represents a target probability distribution. Some of the sampling methods include the popular Gibbs sampler and the Metropolis-Hastings algorithm.

Misalignment: This occurs through the sampling design. For instance, a mobile monitoring station takes environmental data at different locations at different time points. A fixed monitoring station is moved from its current location at one time point to a new location at a different time point.

Monte Carlo methods: Monte Carlo methods (named after a Morocco city) are mathematic experimentation, which typically uses random numbers to conduct experiments on computers for specific purposes.

Point pattern data: These are spatial data generated from a discrete spatial random process. The key characteristic of the point pattern process is that the spatial domain D is itself random and the elements of the index set D mark the spatial locations of events that constitute the spatial point pattern.

Point-referenced data: These are spatial data generated from a continuous spatial random process. The indices marking each event vary continuously in a spatial domain D, and each index is associated with a pair of geographic coordinates to label the spatial position of the event in D. The geographic coordinates can be expressed in Universal Transverse Mercator (UTM, popular with Global Positioning System (GPS) users) or spherical degrees (latitude and longitude) or any other methods. In three dimensional systems, a third coordinate marks the elevation of the observed points.

R: A free, command-driven statistical computing program that was initially launched by a group of scientists in New Zealand. It has a core development team for the base package, with many others contributing numerous packages to it. It uses the S and S-plus language for programming. The S language was developed from the Bell Laboratory of Lucent Technologies Inc. in the USA. The commercial counterpart of R is the S-plus sold by Insightful Inc. in Seattle, Washington State, USA.

Simple Sequential Inhibition: It is a Poisson process that generates regular point patterns in a spatial domain D by using a rejection sampling process. Its mechanism is as follows:

1) Generate the first site s_1 from a homogeneous Poisson process and place a desired radius r around s_1.
2) Generate a candidate site c from the same homogeneous Poisson process.
3) Determine the distance d between s_1 and c. Reject c if $d < r$ and draw a new candidate. Otherwise, keep it and call it s_2.
4) Continue drawing candidate sites and rejecting them if they fall within the radius of any points that are already established.

5) Stop when no new points can be accepted, or when the random draw reached the pre-determined total number of points.

Stationary: In time series analysis, a strictly stationary random process is the one whose distribution probability does not change with time. A week stationarity requires only the mean does not change with time and the covariance is determined by time differences, but not by time per se. This concept is readily expanded into spatial statistics.

Theme: In ArcView GIS, a theme is a collection of all or a subset of the features of a particular feature class in the data source that it is based on, for instance, all highways in a map. Each theme is added into an ArcView GIS project as a layer, which can be manipulated when required.

Thiessen polygons: Thiessen polygons are also called Voronoi polygons. They are one type of spatial tessellation. In a plane, their boundaries define the area that is closest to each point relative to all other points. They are mathematically formed by perpendicularly bisecting the lines between all points.

Weakly stationary spatial process: If the mean and variance of a spatial process is dependent on the separation vector $\mathbf{d} = |\mathbf{s} - \mathbf{s}'|$ between pairs of spatial locations s, the process is said to be stationary. If the mean and variance of a spatial process depends only on the separation distance $d = \|\mathbf{s} - \mathbf{s}'\|$ between pairs of spatial locations s, the process is regarded as weakly stationary.

WinBUGS : A free Bayesian statistic software that is jointly developed by a team of scientists at the MRC Biostatistics Unit, Institute of Public Health in Cambridge, UK and the Department of Epidemiology & Public Health, Imperial College School of Medicine in London, UK. It is implemented in the Microsoft Windows environment. BUGS is a humorous acronym standing for Bayesian Inference Using Gibbs Sampling. The latest version of WinBUGS is 1.4.1. Efforts to update the program have stopped due to the launch of the OpenBUGS program, which can be implemented in R and other statistical packages, such as MatLab, SAS, STATA etc.

Appendix II

Kalman filter

(References: Meinhold and Singpurwalla 1983, West and Harrison 1997)

Define a univariate dynamic linear model as:

Observation equation: $Y_t = F_t' x_t + V_t$ $\quad V_t \sim N(0, \upsilon_t)$ (A1)

System equation: $x_t = G_t X_{t-1} + W_t$ $\quad W_t \sim N(0, \varpi_t)$ (A2)

Initial information: $(x_t | D_0) \sim N(\hat{X}_0, \Sigma_0)$ (A3)

The estimation procedure of the model is as follows:

At time t_0, choose $\hat{X}_0$ and Σ_0 as our best guesses of the mean and variance of X_0.

At time t-1, our knowledge about the state X_{t-1} is:

$$(X_{t-1} \mid Y_{t-1}) \sim N(\hat{X}_{t-1}, \Sigma_{t-1}) \tag{A4}$$

where $\hat{X}_{t-1}$ and Σ_{t-1} are the expectation and the variance of $(X_{t-1} \mid Y_{t-1})$, respectively.

At time t prior to observing Y_t, our best choice of X_t is governed by the system equation (A2). Since X_{t-1} is described by A4, therefore:

$$(X_t \mid Y_{t-1}) \sim N(G_t \hat{X}_{t-1}, R_t) \tag{A5}$$

where $R_t = G_t \Sigma_{t-1} G_t' + W_t$. This step is equivalent to a forecast of X_t based on our knowledge of X_{t-1}.

At time t after observing Y_t, our knowledge of X_t is updated by correcting our forecast based on information on X_{t-1} in the prior step, i.e.

$$(X_t \mid e_t, Y_{t-1}) \sim N(\hat{X}_t, \Sigma_t) \tag{A6}$$

where

$$e_t = Y_t - \hat{Y}_t = Yt - F_t G_t \hat{X}_{t-1} \tag{A7}$$

$$\hat{X}_t = G_t X_{t-1} + R_t F_t^{'} (V_t + F_t R_t F_t^{'})^{-1} e_t \tag{A8}$$

$$\Sigma_t = R_t + R_t F_t^{'} (V_t + F_t R_t F_t^{'})^{-1} F_t R_t \tag{A9}$$

In A8 and A9,

$R_t F_t^{'} (V_t + F_t R_t F_t^{'})^{-1}$ is called the Kalman gain.

The estimation procedure iteratively continues from A5 to A9 until the final destination.

Appendix III

Conditional distributions

(Reference: Gelman, A, Carlin, JB, Stern, HS and Rubin, DB. 1995. Bayesian data analysis. Chapman & Hall. New York, NY. Pp 28-58).

AIII.1. Preparations

AIII.1.1 Basics of Bayesian statistics

Given data *y*, the joint probability distribution for θ and *y* is a product of the prior distribution p(θ) and the sampling distribution p(*y*|θ):

$$p(\theta, y) = p(\theta)p(y|\theta) \qquad \text{A3.1}$$

Conditional on the known value of *y* by using the Bayes' rule, the posterior density is:

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y \mid \theta)}{p(y)} = \frac{p(\theta)p(y \mid \theta)}{\int p(\theta)p(y \mid \theta)d\theta} \propto p(\theta)p(y|\theta) \qquad \text{A3.2}$$

The core of the Bayesian data analysis is to develop the model p(θ, y) and to find the posterior distribution p(θ|y).

AIII.1.2 Posterior distribution of normally distributed data with a normal prior

If both the prior p(θ) and a datum probability distribution p(*y*|θ) are both normal,

$$p(\theta) = \frac{1}{\sqrt{2\pi s}} \exp(-\frac{1}{2s^2}(\theta - m)^2) \qquad \text{A3.3}$$

and

$$p(y \mid \theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp(-\frac{1}{2\sigma_0^2}(y - \theta)^2) \qquad \text{A3.4}$$

then

$$p(\theta|Y) \propto p(\theta)p(y|\theta)$$

$$\propto \exp[-\frac{1}{2s^2}(\theta-m)^2]\times\exp[-\frac{1}{2\sigma_0^2}(y-\theta)^2]$$

$$=\exp[-\frac{1}{2s^2}(\theta^2-2\theta m+m^2)-\frac{1}{2\sigma_0^2}(y^2-2y\theta+\theta^2)]$$

$$=\exp[-\frac{1}{2s^2\sigma_0^2}[\sigma_0^2\theta^2-2\sigma_0^2\theta m+\sigma_0^2m^2-s^2y^2+2s^2y\theta+s^2\theta^2]]$$

$$=\exp[-\frac{1}{2s^2\sigma_0^2}[\theta^2(\sigma_0^2+s^2)-2\theta(\sigma_0^2m+s^2y)+(\sigma_0^2m^2-s^2y^2)]]$$

$$\propto \exp[-\frac{1}{2s^2\sigma_0^2}[\theta^2(\sigma_0^2+s^2)-2\theta(\sigma_0^2m+s^2y)]]$$

$$=\exp\left[-\frac{1}{2}\left[\theta^2\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)-2\theta\left(\frac{m}{s^2}+\frac{y}{\sigma_0^2}\right)\right]\right]$$

$$=\exp\left[-\frac{1}{2}\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)\left[\frac{\theta^2\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)}{\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)}-\frac{2\theta\left(\frac{m}{s^2}+\frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)}\right]\right]$$

$$=\exp\left[-\frac{1}{2}\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)\left[\theta^2-\frac{2\theta\left(\frac{m}{s^2}+\frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)}\right]\right]$$

$$=\exp\left[-\frac{1}{2}\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)\left[\theta^2-\frac{2\theta\left(\frac{m}{s^2}+\frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)}\right]+\mu-\mu\right]$$

$$=\exp\left[-\frac{1}{2}\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)\left[\theta^2-\frac{2\theta\left(\frac{m}{s^2}+\frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)}\right]+\mu\right]\exp[-\mu]$$

$$\propto \exp\left[-\frac{1}{2}\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)\left[\theta^2-\frac{2\theta\left(\frac{m}{s^2}+\frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)}\right]+\mu\right]$$

$$\text{Let } \mu=\left(\frac{\left(\frac{m}{s^2}+\frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)}\right)^2$$

$$p(\theta \mid y) \propto \exp\left[-\frac{1}{2}\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)\left(\theta-\frac{\left(\frac{m}{s^2}+\frac{y}{\sigma_0^2}\right)}{\left(\frac{1}{s^2}+\frac{1}{\sigma_0^2}\right)}\right)^2\right] \qquad \text{A3.5}$$

$p(\theta|y) \sim N(\mu, \sigma^2)$

where $\frac{1}{\sigma^2}=\frac{1}{s^2}+\frac{1}{\sigma_0^2}$

Given a normal prior distribution with hyperparameters of σ_0^2 and μ_0

$p(\theta) = N(\theta|\mu_0, \sigma_0^2)$

and the normally distributed data **Y** (**Y** $= y_1, \ldots, y_n$), i.e. $y_i|\theta \sim N(\theta, \sigma^2)$, the posterior distribution of θ, given **Y**, is:

$$p(\theta|Y) \propto p(Y|\theta)p(\theta)$$

$$= p(\theta)\prod_{i=1}^{n} p(y_i \mid \theta)$$

$$\propto \exp(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2)\prod_{i=1}^{n}\exp(-\frac{1}{2\sigma^2}(y_i-\theta)^2)$$

$$\propto \exp(-\frac{1}{2}[\frac{1}{\tau_0^2}(\theta-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2])$$

With some algebraic manipulations as shown above, the above equation becomes:

$$p(\theta|y_1,\ldots,y_n) = p(\theta|\bar{y}) = N(\theta|\mu_n, \tau_n^2) \quad \text{A3.6}$$

where

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0+\frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}} \text{ and } \frac{1}{\tau_n^2}=\frac{1}{\tau_0^2}+\frac{n}{\sigma^2} \quad \text{A.3.7}$$

AIII.1.3 Gamma, inverse-Gamma, and scaled-inverse-χ^2 distributions

The Gamma distribution takes the form:

$$p(x)=\frac{\theta^{-k}}{\Gamma(k)}x^{k-1}\exp(\frac{-x}{\theta}) \quad \text{A3.8}$$

If x ~ Gamma(k, θ), then $y=\frac{1}{x}$ ~ Inv-Gamma(α, β). The inverse-Gamma distribution (with x > 0) then is:

$$p(y)=\frac{\beta^{\alpha}}{\Gamma(k)}y^{-(\alpha+1)}\exp(\frac{-\beta}{y}) \quad \text{A3.9}$$

With shape parameter α (=k) and scale parameter β (= θ^{-1}) as derived below:

$$f_Y(y)=f_x(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right|$$

$$=\frac{1}{\theta^k\Gamma(k)}\left(\frac{1}{y}\right)^{k-1}\exp\left(\frac{-1}{\theta y}\right)\frac{1}{y^2}$$

$$=\frac{1}{\theta^k\Gamma(k)}\left(\frac{1}{y}\right)^{k+1}\exp\left(\frac{-1}{\theta y}\right)$$

$$=\frac{\theta^{-k}}{\Gamma(k)}y^{-(k+1)}\exp\left(-\frac{\theta^{-1}}{y}\right) \qquad \text{A3.10}$$

For $p(y|\theta, \sigma^2)$ with θ known and σ^2 unknown, the likelihood for a vector **Y** of n identically, independently distributed observations is

$$P(Y|\sigma^2) \propto \sigma^{-n}\exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right) = (\sigma^2)^{-\frac{n}{2}}\exp\left(\frac{-ns}{2\sigma^2}\right) \qquad \text{A3.11}$$

where $s=\frac{1}{n}\sum_{i=1}^{n}(y_i-\theta)^2$ is a sufficient statistic.

Given an inverse-gamma conjugate prior density with hyperparameters α and β,

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)}\exp\left(\frac{-\beta}{\sigma^2}\right) \qquad \text{A3.12}$$

The posterior density for σ^2 is

$$p(\sigma^2|y) \propto p(\sigma^2)p(y|\sigma^2)$$

$$\propto (\sigma^2)^{-(\alpha+1)}\exp\left(\frac{-\beta}{\sigma^2}\right)(\sigma^2)^{-\frac{n}{2}}\exp\left(\frac{-ns}{2\sigma^2}\right)$$

$$\propto (\sigma^2)^{-(\alpha+\frac{n}{2}+1)} \exp(-\frac{1}{\sigma^2}(\beta+\frac{ns}{2})) \qquad \text{A3.13}$$

i.e.

$$p(\sigma^2|y) \sim \text{Inv-Gamma}(\frac{n}{2}+a, \frac{ns}{2}+\beta) \qquad \text{A3.14}$$

AIII.2 Conditional distribution for μ_t, λ_ε and λ_v

Follow the same logic as above for the normally distributed data and re-arrange Eqn 3.3a

$$y_{s,t} = \mu_{s,t} + \sum_{i=1}^{M} k_t(\omega_i - s)x(\omega_i, t) + \varepsilon_{s,t}$$

to obtain

$$\mu_{s,t} = y_{s,t} - \sum_{i=1}^{M} k_t(s)x_{i,t}$$

Define

$$a_t = \sum_{s=1}^{N_t}(y_{s,t} - k_t(s,.)x_t) \text{ (equivalent to } n\bar{y})$$

where $\mathbf{k}_t(s,\cdot)$ is the sth row of the smoothing kernel matrix $\mathbf{K}_t$. According to Eqn A3.7, the full conditional distribution of μ_t is

$$\mu_t \,|\, - \sim N(\frac{\frac{a_t}{\lambda_\varepsilon}+\frac{m_\mu}{c_\mu}}{\frac{N_t}{\lambda_\varepsilon}+\frac{1}{c_\mu}}, \frac{N_t}{\lambda_\varepsilon}+\frac{1}{c_\mu})$$

To find λ_ε, define

$$SS_\varepsilon = \sum_{t=1}^{T} (\mathbf{Y}_t - \mu\mathbf{1} - \mathbf{K}x_t)'(\mathbf{Y}_t - \mu\mathbf{1} - \mathbf{K}x_t)$$

According to A3.14, the full conditional distribution of λ_ε is

$$\lambda_\varepsilon \,|\, - \sim IG((\frac{NT}{2} + \gamma_\varepsilon, \frac{SS\varepsilon}{2} + \delta_\varepsilon)$$

Similarly, define

$$SS_v = \sum_{t=1}^{T} (x_t - \alpha x_t)'(x_t - \alpha x_t)$$

The full conditional distribution of λ_v is

$$\lambda_\varepsilon \,|\, - \sim IG((\frac{MT}{2} + \gamma_v, \frac{SSv}{2} + \delta_v)$$

The full conditional distributions of the x_t's are found sequentially using the Kalman filter (Meinhold and Singpurwalla 1986, West and Harrison 1997).

Appendix IV

Computing and graphing programs

Note: # stands for comments. Anything following it would not be executed in R or WinBUGS.

# 'Root.directory' is the working directory on the C or any other drive.

AIV 1. WinBUGS codes for the main spatiotemporal model

```
Model { # note T=12+1 in my research.

# The observation Equations

for (t in 2:T) {
  for (s in offset[t]:(offset[t+1]-1)) {
 m.y[s] <- inprod(K[s, ], X[, t]) + mu[t]
   y[s] ~ dnorm(m.y[s], tau.y)   }    }

# The state Eqn

for (t in 2:T) {
   for (m in 1:M) {
   X[m, t] ~ dnorm(X[m, (t-1)],  tau.x) }     }

#  The priors

 for (i in 1:M) { m.X[i] <- 0;
 C.inv[ i] <- 0.05;
 X[i, 1] ~ dnorm(m.X[i], C.inv[i])}
for (t in 2:T) {mu[t] ~    dnorm(0, 100)}
tau.y ~ dgamma(0.1, 0.1);
sig2.y<-1/tau.y;
tau.x ~ dgamma(.1, .1);
sig2.x<-1/tau.x;
  }
```

Initials (a *separate dat file*)

Data (a *separate dat file*)

AIV. 2. R codes for data preparation and analysis

1. Codes to produce the supporting sites

```
sppt.site.fn<-function(data, sd, plot.it = F) {
# data: The original monitoring dataset; sd: the standard deviation of the Gaussian
# kernel. A buffering zone of the width of sd is added to surround the study region
# vertical distance between rows
    ht <- (sd/2) * tan(pi/3)
# distance between rows
    sw <- c(min(data$x) - sd, min(data$y) - sd)
    ne <- c(max(data$x) + sd, max(data$y) + sd)
# number of columns of supporting sites
    n <- 1 + round((ne[1] - sw[1])/sd)
# number of rows of supporting sites
    m <- round((ne[2] - sw[2])/ht)
    if (!is.integer(ne[2]-sw[2]/ht - m)) m<-m+1
# margines of x, y coordinates at the sw corner of the study region
    mn.x <- sw[1] - sd/2
    mn.y <- sw[2] - sd/2
    # initialize the first two rows of supporting sites
    rw1.x <- numeric(length = n)
    rw1.y <- numeric(length = n)
    rw2.x <- numeric(length = n - 1)
    rw2.y <- numeric(length = n - 1)
# row 1
    rw1.x[1] <- mn.x
    rw1.y[1:n] <- mn.y
    for(i in 2:n) {
    rw1.x[i] <- rw1.x[(i - 1)] + sd }
# row 2
    rw2.x[1] <- sd/2 + mn.x
    rw2.y[1:(n-1)] <- ht + mn.y
    for(i in 2:(n - 1)) {
    rw2.x[i] <- rw2.x[(i - 1)] + sd }
xx <- append(rep(c(rw1.x, rw2.x), 0.5 * m + 1), rw1.x)
    yy <- append(rw1.y, rw2.y)
    yy.n<-length(yy)
    for(i in 1:(0.5 * m)) {
    yy <- append(yy, c(rw1.y, rw2.y) + ht * (i * 2)) }
    yy.n<-length(yy)
    yy <- append(yy, rw1.y + (m + 1) * ht)
    n <- length(yy)
    nn <- as.integer(seq(1, n, by = 1))
    sppt.site <- data.frame(cbind(nn, xx, yy))
    if(sppt.site[yy.n, 3]==sppt.site[yy.n+1, 3]) sppt.site<-sppt.site[1:yy.n, ]
    dimnames(sppt.site)[[2]] <- c("sppt.site", "x", "y")
    if(plot.it) {
    plot(sppt.site[, 2], sppt.site[, 3], type = "n", xlab = "X", ylab = "Y")
    text(sppt.site[, 2], sppt.site[, 3], labels = sppt.site[, 1], cex = 0.5) }
    sppt.site}
```

2. Codes to produce the bivariate Gaussian kernel

```
ker.fn <-function (sppt.data,  site.data, sd) {
     # sppt.data is a data frame holding the x, y coordinates of the supporting sites
```

```
    # that are laid on a grid covering the study region.
    # site.data is a list of data frames. Each data frame holds the time, site_id, and
    # x,y coordinates of the sites at time t in columns.
    # sd is the standard deviation of the Gaussian kernel.
    sx<-sppt.data$x;          sy<-sppt.data$y;
    ns <- length(sx);                # number of supporting sites
    T <- length(site.data);          # total number of monitoring time points
    nm<-numeric(length=T);           # number of monitoring sites at time t
    # Create the Gaussian kernel
    ker<-list(length=T);
    for (t in 1:T){
    nm[t] <-length(site.data[[c(t,3)]]);
    dx <-matrix(NA, nrow= nm[t], ncol=ns);
    dy <-matrix(NA, nrow=nm[t], ncol=ns);
    ker.t <-matrix(NA, nrow=nm[t], ncol=ns);
    for (i in 1:nm[t]){
    dx[i, ]<-dnorm(sx, mean=site.data[[c(t,3,i)]], sd=sd);
    dy[i, ]<-dnorm(sy, mean=site.data[[c(t,4,i)]], sd=sd);
    }
    ker.t <-dx*dy;
    #  scale the sum of the squared kernel at each monitoring site to 1
    d<-sqrt(rowSums(ker.t *ker.t));
    for (i in 1:nm[t]){ker.t [i,]<-ker.t [i,]/d[i]};
    dimnames(ker.t) <- NULL; # Remove dim names
    ker[[t]]<-t(ker.t); }
return(ker) }
```

3. R codes to pool the kernel matrices for the WinBUGS

```
ker.pooled<-rbind(ker[[3]], ker[[4]])        # use only the 12 measured months.
for(i in 5:14){ker.pooled<-rbind(ker.pooled, ker[[i]])}
dimnames(ker.pooled)<-NULL
```

4. R codes to produce the offset vector for the WinBUGS

```
# N is a vector holding the number of sites measured at each time period.
# T is the total time point (i.e. 12 in my research)
offset<-numeric(length=(T+1));
offset[1]<-NA;
offset[2]<-1;
for (i in 3:T) {offset[i]<-offset[2]+sum(N[(3-1):(i-1)])};
offset[T+1]<-sum(N[!is.na(N)])+1;
 (sum(N[!is.na(N)]); offset
# Check if the largest entry of the offset vector equals total number of sites +1
# It is very important to get this vector right).
```

5. R codes to produce the data list for the WinBUGS

```
data.list<-list(T, M, offset, y, ker)
 # T = total time points +1, M = number of supporting sites.
 # offset is the offset vector created above.
# y is the log_so2 data vector
# ker is the pooled kernel matrix.
names(data.list)<-c('T', 'M', 'offset', 'y', ‘K’)
writeDatafileR(data.list, 'Root.directory/data.txt')
# download the above writeDatafileR function from the Internet.
# WinBUGS accept capital ‘E’, but not small ‘e’, in scientific notation, and
# no spaces between any numbers preceding the ‘E’.
# All these can be edited out by in MS Word by using  the ‘replace’
# function in the ‘edit’ pull down menu.
# The Initials list can be created in a similar way according to the parameters
# to be estimated in the WinBUGS model.
```

6. R codes to produce the semivariograms and the graph

```
geo.dat <- as.geodata(cbind(data$x, data$y, data$log_so2))
variog.dat <- variog(coords = geo.dat$coords, data = geo.dat$data,
 trend='2nd', method='matern', kappa=0.45, type='modulus')
cov.dat <- variofit(variog.dat, ini.cov.pars=c(0.25, 120), cov.model = "matern",
kappa=0.45)
plot(variog.dat, xlab='Distance (km)',ylab='Semivariogram')
lines(cov.dat)
legend(x='topleft', legend=c('Jun 2001'), bty='n')          # replace the legend
text(200, 0.02, labels=expression(~tau^2))
text(375, 0.02, labels=expression(~sigma^2))
text(545, 0.02, labels=expression(~phi))
text(265, 0.02, labels='= 0.20')     # replace the value following the labels =
text(435, 0.02, labels='= 0.15')     # replace the value following the labels =
text(605, 0.02, labels='= 230')      # replace the value following the labels =
```

7. R code for the cluster analysis and plotting the results

```
tree.dat<-hclust(dist(poly.attributes.dat), method = "ward")
plot(tree.dat, ylab = "Height", cex=0.3)
group.dat<-cutree(tree.dat, k = 4) # replace the k value
```

AIV.3. SAS codes

(These were provided by Dr. Igor Burstyn)

1 The Linear mixed effects model

```
Proc import out=my.data
datafile="root.directory:\my.Excel.data file"  */change file name here
dbms=excel replace;
sheet="my.Excel.data$";
getnames=yes;
mixed=no;
scantext=yes;
useddate=yes;
scantime=yes;
```

```
run;

proc mixed data=my.data covtest;
class location time;
model log_so2=/outp=predicted_value;
random int/subject=location;
random int/subject=time;
run;
```

2 Temporal semivariogram

```
libname Work 'c:\programs\sas\libraries\work';
data v;
set data.file.name;        */change file name here
z=0;
run;

proc variogram data=v outv=outv;
compute lagd=1 maxlag=12 robust;
coordinates xc=time /*month from 1 to 12*/ yc=z /*z=0=constant*/;
var logso2;
run;

title 'OUTVAR= Data Set Showing Temporal Variogram Results for ln(SO2)';
data outv; set outv;
if varname='log_so2' then name='SO2' ; proc sort data=outv; by name; proc print
data=outv label;
var lag count distance variog rvario;
by name;
run;

data outv2; set outv;
```

```
vari=variog; type = 'regular'; output;
vari=rvario; type = 'robust'; output;
run;

title 'Standard and Robust Semivariogram for ln(SO2)';
proc sort data=outv2;
by name;
proc gplot data=outv2;
plot vari*distance=type / frame vaxis=axis2
haxis=axis1;
symbol1 i=join l=1 v=star;
symbol2 i=join l=1 v=square;
axis1 minor=none
label=(c=black 'Lag Distance (month)') /* offset=(3,3) */;
axis2 order=(0 to 1 by 0.1) minor=none
label=(angle=90 rotate=0 c=black 'Variogram')
/* offset=(3,3) */;
run;
```

3. Random effects model comparing dendrogram cut effects on variances

```
* Codes for the four class cut. Replace variable name for the seven class cut
proc mixed data=data method=ml covtest;
class poly_grp4 poly_id;
model y=;
random poly_grp4 poly_id(poly_grp4);
run;
```

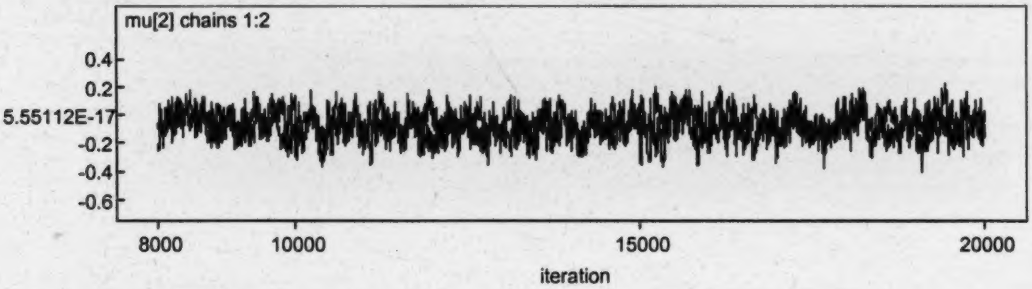
Appendix V

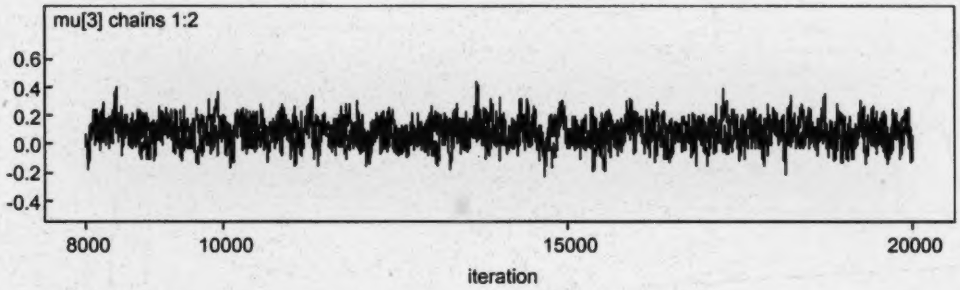
Coordinates of the 64 supporting sites used for the spatiotemporal modeling

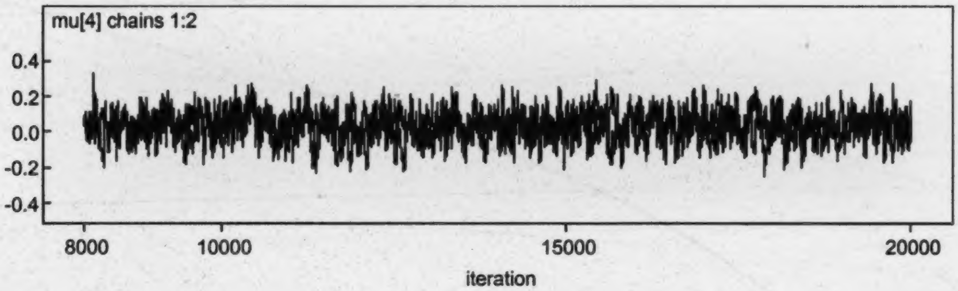
No.	Site_x	Site_y		No.	Site_x	Site_y
1	7154.709	11369.27		33	7954.709	12408.5
2	7354.709	11369.27		34	8154.709	12408.5
3	7554.709	11369.27		35	7254.709	12581.7
4	6854.709	11542.47		36	7454.709	12581.7
5	7054.709	11542.47		37	7654.709	12581.7
6	7254.709	11542.47		38	7854.709	12581.7
7	7454.709	11542.47		39	8054.709	12581.7
8	7654.709	11542.47		40	8254.709	12581.7
9	6954.709	11715.68		41	7354.709	12754.91
10	7154.709	11715.68		42	7554.709	12754.91
11	7354.709	11715.68		43	7754.709	12754.91
12	7554.709	11715.68		44	7954.709	12754.91
13	7754.709	11715.68		45	8154.709	12754.91
14	7054.709	11888.88		46	8354.709	12754.91
15	7254.709	11888.88		47	7254.709	12928.11
16	7454.709	11888.88		48	7454.709	12928.11
17	7654.709	11888.88		49	7654.709	12928.11
18	7854.709	11888.88		50	7854.709	12928.11
19	7154.709	12062.09		51	8054.709	12928.11
20	7354.709	12062.09		52	8254.709	12928.11
21	7554.709	12062.09		53	8454.709	12928.11
22	7754.709	12062.09		54	7154.709	13101.32
23	7954.709	12062.09		55	7354.709	13101.32
24	7254.709	12235.29		56	7554.709	13101.32
25	7454.709	12235.29		57	7754.709	13101.32
26	7654.709	12235.29		58	7254.709	13274.52
27	7854.709	12235.29		59	7454.709	13274.52
28	8054.709	12235.29		60	7654.709	13274.52
29	7154.709	12408.5		61	7854.709	13274.52
30	7354.709	12408.5		62	7354.709	13447.73
31	7554.709	12408.5		63	7554.709	13447.73
32	7754.709	12408.5		64	7754.709	13447.73

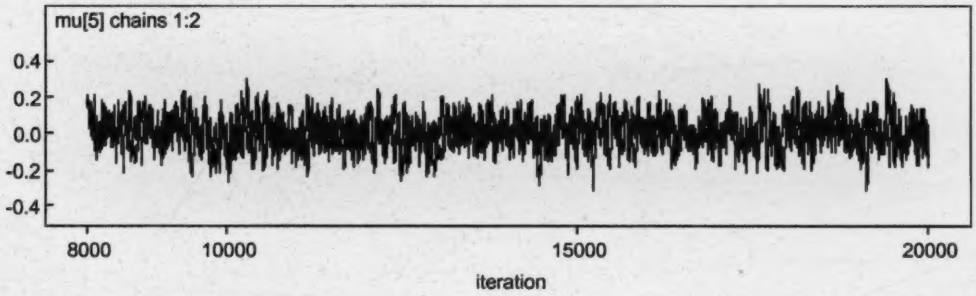
Appendix VI

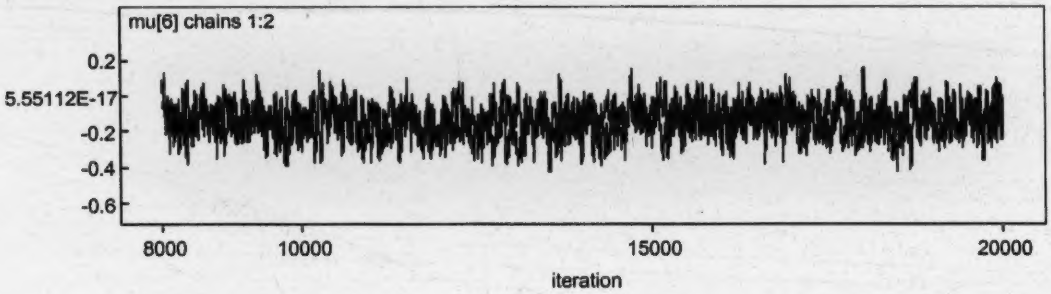
History series of the spatiotemporal model parameters

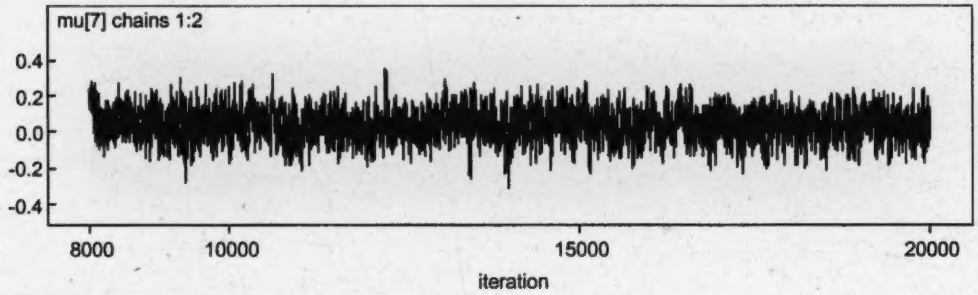


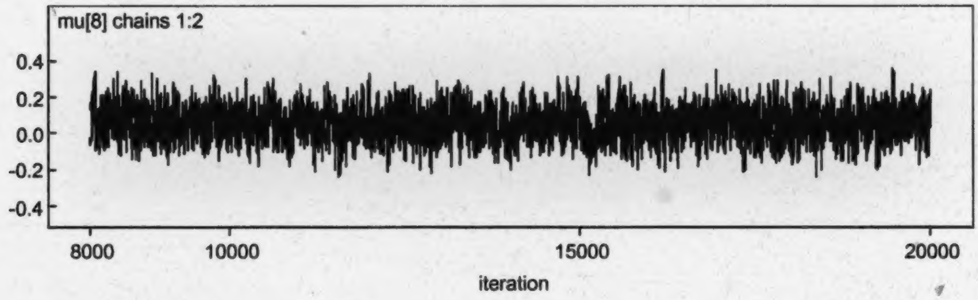


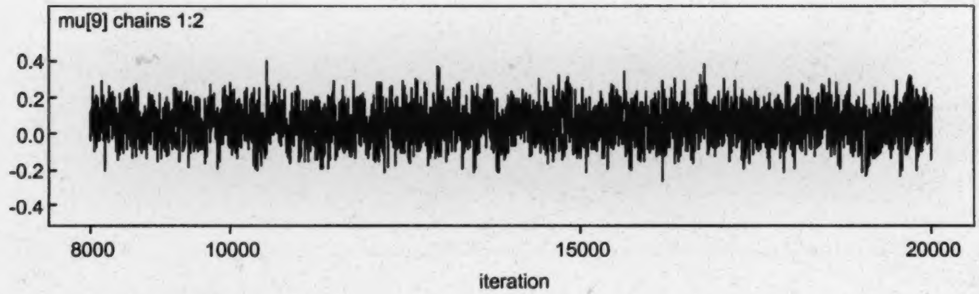


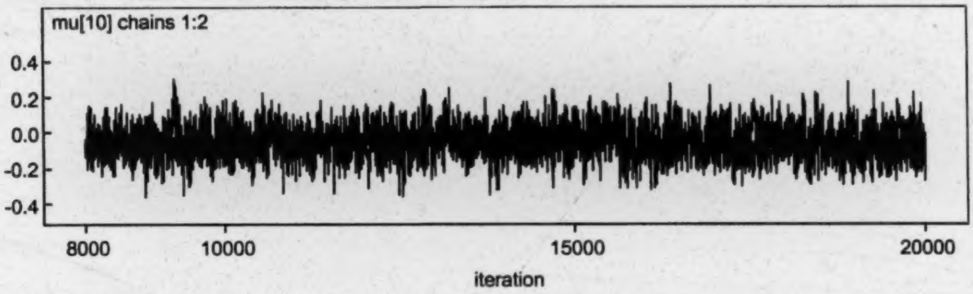


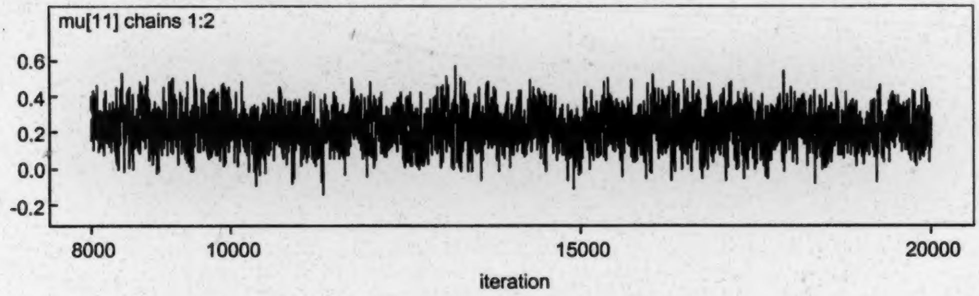

mu[6] chains 1:2
0.2
5.55112E-17
-0.2
-0.4
-0.6
8000
10000
15000
20000
iteration

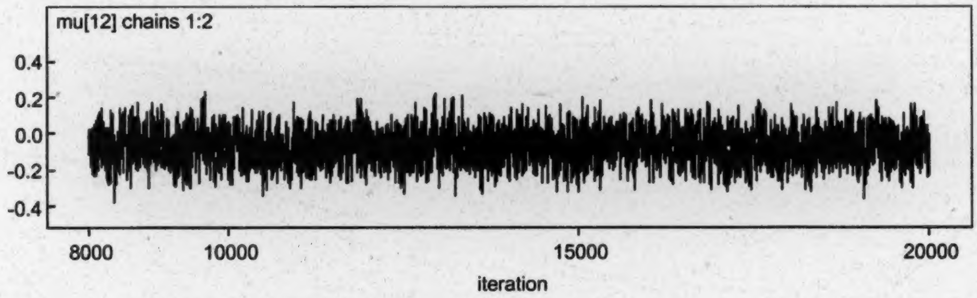

mu[7] chains 1:2
0.4
0.2
0.0
-0.2
-0.4
8000
10000
15000
20000
iteration

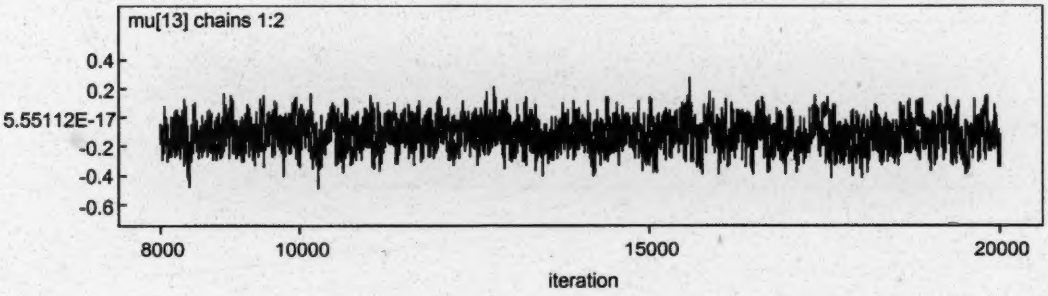

mu[8] chains 1:2
0.4
0.2
0.0
-0.2
-0.4
8000
10000
15000
20000
iteration

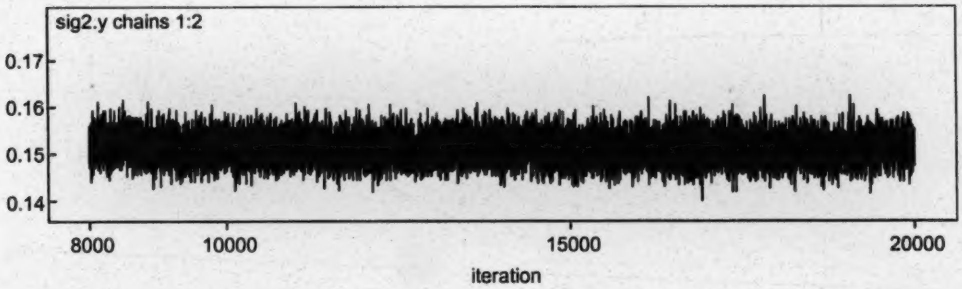

mu[9] chains 1:2
0.4
0.2
0.0
-0.2
-0.4
8000
10000
15000
20000
iteration

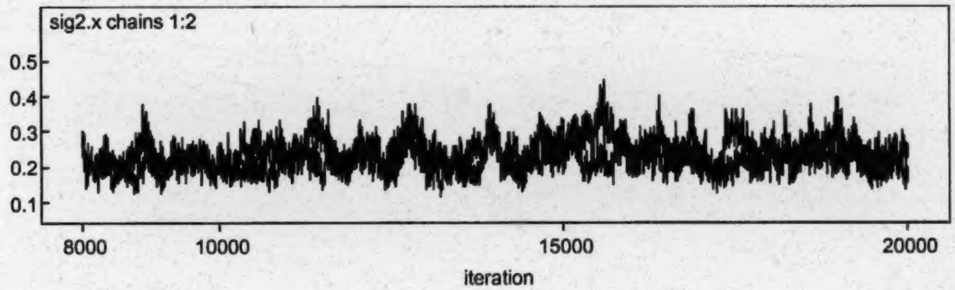

mu[10] chains 1:2
0.4
0.2
0.0
-0.2
-0.4
8000
10000
15000
20000
iteration


mu[11] chains 1:2
0.6
0.4
0.2
0.0
-0.2
8000
10000
15000
20000
iteration


mu[12] chains 1:2
0.4
0.2
0.0
-0.2
-0.4
8000
10000
15000
20000
iteration


mu[13] chains 1:2
0.4
0.2
5.55112E-17
-0.2
-0.4
-0.6
8000
10000
15000
20000
iteration





Notes: mu[2] ~ mu[13] are equivalent to μ_1 to μ_{12}, and sig2.x and sig2.y are equivalent to λ_υ and λ_ε in the spatiotemporal model.

Appendix VII

SAS modeling results

It was proposed that a comparison be made between this spatiotemporal model and a previously established random-effect linear model estimated through the Proc Mixed procedure using the Best Linear Unbiased Estimation method in SAS (Dr. Burstyn, pers. comm.). The SAS model took the form:

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where **Y** is a vector of observed $\ln(SO_2)$, **Z** a matrix of monitoring sites and time points, $\boldsymbol{\beta}$ a vector of unknown random-effects parameters, $\boldsymbol{\varepsilon}$ a vector of unknown independent and identically distributed normal (Gaussian) random variables with mean 0 and variance σ^2.

The SAS codes that treat both the monitoring sites and the monitoring time as random-effect are shown in Appendix VI. The model results are listed in Table A5.1, which show a 39% greater of spatial variance than temporal variance. The relationship between the model fitted and the observed $\ln(SO_2)$, as well as the residuals of the model fitted values are shown in Figure A5.1. As typical with any random-effect model, the fit of this specific random-effect model to the SO_2 data is sufficiently well. The residuals are generally normally distributed, though slightly left-skewed. This model also handled the problematic observations well. These results are broadly comparable to those of the spatiotemporal model as shown in Figure 3.7, with the spatiotemporal model producing a better residual distribution. The fitted $\ln(SO_2)$ values of the two models are linearly correlated (Figure A5.2), with an adjusted $R^2 = 0.75$.

This SAS model has three key assumptions, i.e. normality, homoscedasticity and independence. Unfortunately, as the SO_2 data arose from a spatiotemporal random process, measurements of the ambient SO_2 concentrations were both temporally and spatially correlated. The value of individual SO_2 measurements is random, but not the location where the value is obtained.

One weakness of this SAS model emerges if one wants to use it for predicting the ambient SO_2 concentrations for new locations in the study region. The entire model needs to be re-computed and the results could be variable from time to time when a new prediction is performed. In contrast, there is no such a requirement for re-computation of the spatiotemporal model per se in this usage.

Table A5.1. Covariance Parameter Estimates of the SAS model

Covariance Parameter	Subject	Estimate	Standard Error	Z Value	Pr Z
Intercept	Location	0.25	0.01	22.19	<.0001
Intercept	Time	0.18	0.08	2.34	0.0096
Residual		0.14	0.00	67.45	<.0001

Fit Statistics

-2 Res Log Likelihood	12403.3
AIC (smaller is better)	12409.3
AICC (smaller is better)	12409.3
BIC (smaller is better)	12403.3

Figure A5.1. Positive linear relationship between the SAS modeled and the measured ln(SO_2) (a) and the residuals of the SAS modeled ln(SO_2) (b). The red dots in panel (a) are the problematic observations as shown in Figure 3.7.

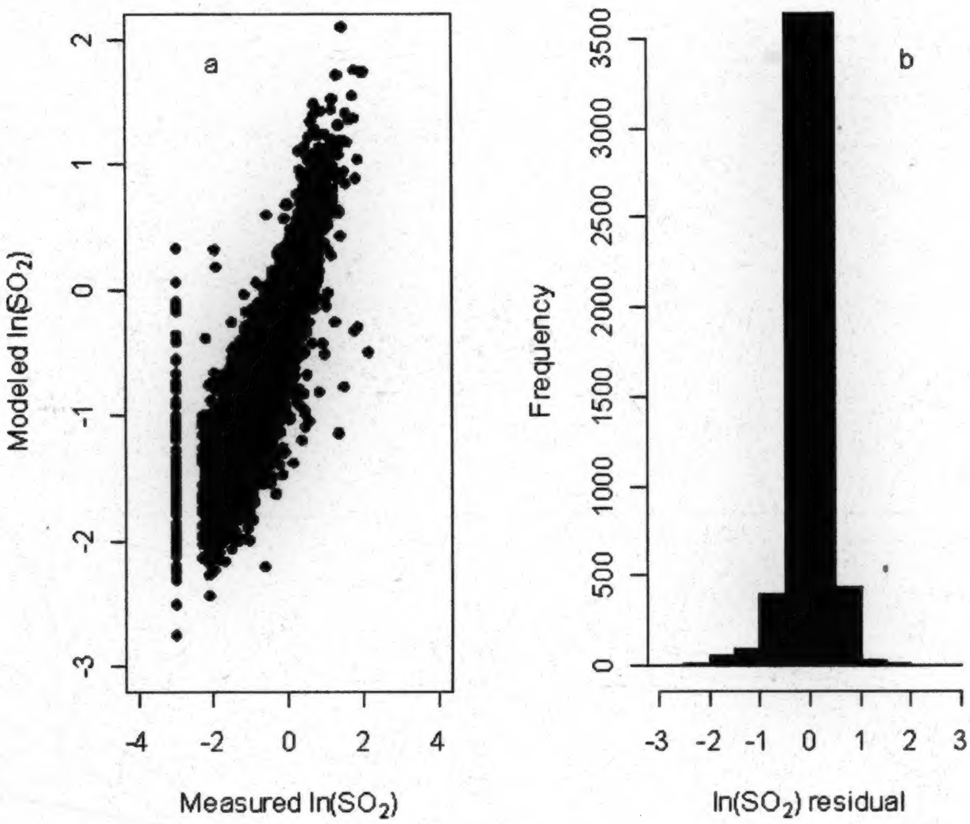


Figure A5.2. Linear correlation between the ln(SO_2) fitted by SAS (SAS fit) and by the spatiotemporal mode (ST fit)